International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-2, Issue-11, November 2016 Pages 62-75

Automatic Extraction of Topics from Documents: Five Probabilistic Topic Model Tests

Sandra Jhean-Larose, Nicolas Leveau, Guy Denhière , Ba-Linh Nguyen

***Abstract*-In this paper, we test the capability of the Topic[1] model to extract topics from documents (Griffiths &Steyvers, 2003, 2004; Griffiths, Steyvers&Tenenbaum, 2007). After presenting the mathematical aspects of the model and demonstrating its behavior on a small corpus, we attempt to falsify the model by manipulating (i) the size and similarities between the sub-corpora, (ii) the relative weight of sub-corpora,and (iii) the permeability to the scope and nature of contexts added to a fixed corpus. The model successfully passed our five tests, demonstrating that first, extracted topics were relevant and congruent to the content of the corpus, and second, that their probability appropriately reflected the relative weight of sub-corpora.**

***Index Terms*- Comprehension, Documents, Probabilistic Models, Semantic memory, Topics.**

I. Introduction

The conceptualizations of reading comprehension and memory models are closely interlinked (Denhière, Lemaire, Bellissens&Jhean-Larose, 2004; Tiberghien, 1997). The interaction between memory and comprehension is all the more significant as a majority of the most influential models of comprehension postulate that word meanings (or concepts) are not stored as "ready-made" units in a mental lexicon but are generated and contextualized in the working memory from memory traces activated in long-term memory. Drawing on the research undertaken by Barclay, Bransford, Franks, McCarrell&Nitsch (1974), numerous studies support the assertion of an emergent nature of the contextual construction of meaning: while in the "furniture" context, pianos are "heavy", in the "Rubinstein" context they are "musical"(Blanc &Brouillet, 2003).

The changes in text comprehension models proposed by Kintsch since 1978 can help to illustrate the interdependence between memory and comprehension. For instance, the Construction-Integration model he developed (Kintsch, 1988) differed from the Kintsch and van Dijk model proposed in 1978, leading to a shift from a top-down to a bottom-up model. This involved considering the main characteristics of associative memory models proposed in particular by Hintzman (1984) and Murdock (1982), and the introduction of relaxation algorithms proposed by connectionist models such as PDP (Rumelhart& McClelland, 1986) to account for activation propagation in the network and the deactivation or forgetting of irrelevant meanings of polysemic words. Similarly, the construction-integration model proposed in 1998 drew on Latent Semantic Analysis (Landauer&Dumais, 1997) to represent semantic information in long-term memory and demonstrate how expertise and previous knowledge influence understanding.

The key contribution of Latent Semantic Analysis (LSA) (Landauer&Dumais, 1997) or Hyperspace Analogue to Language (HAL) (Burgess &Lundt, 1997) is their ability to represent the content of semantic memory through the use of large text corpora (see Bellissens, Thérouanne&Denhière, 2004). In LSA, the singular value decomposition of the matrix containing word counts per paragraph (or document) is followed by a reduction of the number of dimensions of the resulting matrix. A word's meaning is expressed by a vector with n (± 300) dimensions (Landauer, McNamara, Dennis &Kintsch, 2007),and the value of the cosine of the angle between two vectors determines the semantic similarity between two words or statements (Denhière, Lemaire, Bellissens&Jhean-Larose, 2007). These representations of human, child or adult memories can then be automatically connected to a text comprehension model such as the Construction-Integration model to simulate how memory and previous knowledge intervene in reading comprehension (Lemaire, Denhière, Bellissens&Jhean-Larose, 2006), learning, and knowledge acquisition (Lifchitz, Jhean-Larose &Denhière, 2009; Jhean-Larose, Leclercq, Diaz, Denhière, Bouchon-Meunier, 2010).

One of the greatest difficulties encountered by this type of model lies in the contextualization of meaning that develops in the working memory when context-based meaning emerges (Denhière&Tapiero, 1996). Indeed, representations stored in long-term memory are "decontextualized" and LSA provides only one vector that combines all the meanings of a word: thus the vector "game" might refer to the free physical or mental activity of child or adult, activities based on rules that give rise to gains or losses, an actor's or comedian's performance in a theatre or a film, the movement of a body or mechanism… without forgetting the metaphorical uses of the word. Kintsch (2001, 2008) proposed the predication algorithm to contextualize meaning and activate only common neighbors of a proposition's predicate and arguments.

Sandra Jhean-Larose, Orléans University, Education, Discourse and Learning Laboratory, Paris-Descartes University,45, Rue des Saints Pères,75005,Paris, France.
Nicolas Leveau, EPHE, University Paris VIII,CHArt, Human and Artificial Cognition Laboratory, 41 Rue Gay-Lussac, 75005 Paris,
Guy Denhière, CNRS-EPHE, University Paris VIII,CHArt, Human and Artificial Cognition Laboratory, 41 Rue Gay-Lussac, 75005 Paris.,
Ba-Linh Nguyen, EPHE, University Paris VIII,CHArt, Human and Artificial Cognition Laboratory, 41 Rue Gay-Lussac, 75005 Paris,

[1]Henceforth, we will use capitals to characterise the Topic model in order to differentiate it from topics extracted from documents.

As Kintsch and Mangalath (2011) have argued, the use of predication algorithms as a process of activation propagation within a network has some drawbacks associated with the number of neighbors to activate and the activation threshold to use. One of the ways in which these problems can be completely avoided is to abandon systems such as LSA, which only provide decontextualized abstractions of meaning, and opt instead for probabilistic models in which, by definition, meaning is context-sensitive. Referring back to the example of "game" in a probabilistic model, the term "game" can be assigned to different "topics" in documents relating to "child development", "sport", "theatre" and "auto mechanics."

In probabilistic Topic models, the semantic properties of words and documents are expressed in terms of probability distribution over latent variables known as "topics" (Blei, Ng & Jordan, 2003; Griffiths &Steyvers, 2003, 2004; Griffiths, Steyvers&Tenenbaum, 2007; Hofmann, 1999). These models assume that documents are mixtures of topics, i.e. a probability distribution over a set of topics. Each topic thus contains all the words of a corpus to which a probability is assigned (Denhière, Leveau&Jhean-Larose, 2010). Some words thus represent a given topic more accurately than others, and all words, regardless of whether they are polysemic or homophones/homographs or not, are assigned to several themes according to varying probabilities.

To extract topics and the associated probability distributions, the co-occurrence matrix from the learning corpus is split into two matrices Φ and Θ using the Latent Dirichlet Allocation – LDA – (Blei, Ng, & Jordan, 2003). Φ indicates theimportant words for a given topic, while Θ indicatesthe important topics for a document corpus. The semantic similarity between two words w_1 and w_2 is thus estimated using conditional probability, i.e. $P(w_1 \mid w_2)$, which allows reasonable comparisons of the probability of w_1 across choices of w_2. The mathematical aspects of the model and how it works (1. Generation of co-occurrence matrix; 2. Choice of the number of topics; and 3. Extraction of probability distributions) on a reduced corpus comprising 12 sentences are presented in the appendix.

In sum, there are three major differences between Topic models and LSA. Topic models are able to: (i) propose semantic representations using exploitable latent variables, i.e. topics, (ii) account for asymmetrical relationships between concepts, and (iii) contextualize the long-term working memory (Kintsch&Mangalath, 2011).

II. Objective:The falsification of the Topic Model

The experiments presented below sought to test the capacity of the model to automatically extract relevant topics from a corpus subjected to explicit construction constraints. Our objective was to attempt the falsification of the Topic model and determine its limits. To this end, we built a corpus of French documents and controlled and/or manipulated the following characteristics:
- quantitative (number of sub-corpora, number of documents, number of occurrences, number of different words),
- qualitative (nature and similarity of information).

The following dependent variables were considered:
- the relevance of extracted topics in the light of the properties of the sub-corpora. Pertinence was estimated by the 12 words with the highest probability extracted from these topics.
- the probability of extracted topics in the light of the relative weight of the sub-corpora (number of documents) constituting the corpus considered.

Specifically, we addressed the following questions:

Question 1: Relevance of extracted topics. Do the four topics extracted by the Topic model accurately represent (nature and probability) the four sub-corpora experimentally combined into a single corpus? Does the semantic similarity between the four sub-corpora affect this extraction? If so, does increasing the size of the corpus compensate for these effects?

Question 2: Relative weight of the sub-corpora. If the size of the corpus is kept constant, to what extent does the relative weight of the four sub-corpora that constitute the corpus (estimated by the number of documents) affect the nature and probability of the four topics extracted?

Question 3: Permeability to the nature and scope of the context. Doesaddinga corpus with a different nature ("Le Monde 99", "Encyclopedia", "Literature") but the same size (number of documents) to a given corpus (for instance, "Sports")modify the number, nature and probability of "Sports" topics in the six topics extracted?

Does adding a different and larger corpus to a given corpus (for instance, "sports") modify the number, nature and probability of "sports" topics in the six topics extracted?

III. PROCEDURE

A. Test 1: Relevance of extracted topics in relation to the nature of sub-corpora

Test 1a: Influence of the similarity between sub-corpora

We built three corpora from four sub-corpora each comprising 40 documents with an equal number of occurrences. Three levels of similarity between constituent sub-corpora were established (see Table 1):

- Non-similarity ("Iron age", "Breathing", "Cycling", "Video games"),
- 2*2 similarity ("Breathing – Digestion" and "Football – Rugby"),
- Maximum similarity between the four sub-corpora ("Football, Rugby, Cycling, Boxing").

Table 1: Characteristics of the three corpora (4x40 documents) and four sub-corpora used in test 1a[2].

Corpus	Sub-corpus	Documents	Words	Different Words
Non	"Iron age"	40	1914	985
Similarity	"Breathing"	40	2048	999
4*40	"Cycling"	40	2081	978
	"Videogames"	40	2008	1117
2*2	"Breathing"	40	2048	999
Similarity	"Digestion"	40	2161	875
4*40	"Football"	40	1772	935
	"Rugby"	40	2147	1140
Maximum	"Boxing"	40	1967	912
Similarity	"Cycling"	40	1649	783
4*40	"Football"	40	1683	872
	"Rugby"	40	1960	1019

Four topics were extracted. The nature of the first 12 words[3] and the probability of each extracted topic were considered as dependent variables.

Results

Table 2 presents – for the three corpora in order of increasing similarity – the 12 most probable words from the four topics ranked in order of decreasing probability.

Non-similarity condition: There was a perfect congruence between the extracted topics and the sub-corpora:
- Extracted themes respectively referred to "Breathing", "Iron age", "Video games" and "Cycling", and they corresponded to the nature of the four sub-corpora
- No overlapping of topics was observed between the four topic
- The neighborhood probability of 0.250 for extracted topics was representative of the relative weight of the sub-corpora (25% each).

2*2 similaritycondition:The congruence between the extracted topics and the sub-corpora was acceptable:

- Extracted themes respectively referred to "Rugby", "Digestion", "Breathing" and "Football".
- We observed overlapping across similar topics("Football" and "Rugby" for topics 1 and 4, and "Water" for topics 2 and 3).
- The probability of extracted topics ranged from 0.274 to 0.223.
- The sum of the probabilities of similar topics ("Rugby-Football" and "Digestion-Breathing") was close to 0.500.

Maximum similarity condition: There was no congruence between the extracted topics and the sub-corpora; extracted themes don't reflected one but several sub-corpora: "Boxing and Football" (Topic 1), and "Cycling, Rugby and Football" (Topics 2, 3 and 4).

Conclusion

Comparing the results obtained in the three similarity conditions shows that as similarity across sub-corpora increases, the Topic model experiences increasing difficulty in accurately extracting topics congruent to the sub-corpora content (nature and probability).

Test 1b: Influence of the size of the corpus in the maximum similarity condition

In this test, following the results previously obtained, the similarity effect was examined in more detail: maximum similarity was kept constant and the size of the corpus varied. Two conditions formed respectively of corpora comprising 4x250 and 4x500 documents were added to the previous maximum similarity condition ("Football, Rugby, Cycling, Boxing") (see Table 3).

If the difficulty previously encountered in extracting relevant topics resulted exclusively from the size of the corpus used, then the quality of extraction would improve when the size of the corpus was increased.

Table 3: Characteristics (Number of documents, words, and different words) of the three corpora (4x40, 4x250 and 4x500 documents) used in test 1b.

Corpus	Sub-corpora	Documents	Words	Different words
4x40	"Boxing"	**40**	**1 967**	**912**
	"Cycling"	**40**	**1 649**	**783**
	"Football"	**40**	**1 683**	**872**
	"Rugby"	**40**	**1 960**	**1 019**
4x250	"Boxing"	**250**	**10703**	**4475**
	"Cycling"	**250**	**8492**	**3526**
	"Football"	**250**	**9212**	**3822**
	"Rugby"	**250**	**10572**	**4190**
4x500	"Boxing"	**500**	**26 225**	**5 443**
	"Cycling"	**500**	**21 510**	**4 465**
	"Football"	**500**	**21 770**	**4 980**
	"Rugby"	**500**	**25 029**	**5 453**

As before, four topics were extracted and the 12 most probable words of each topic extracted were considered.

Results

Table 4 shows significant changes in the nature and probability of the four topics extracted by the model. The relevance of the topics increased as the size of the corpus increased.

It should be noted that in the **4x40 condition**, there were no unequivocal topics.
The 4x250 condition: We obtained a "super-topic" with a probability of 0.453 grouping together "Football + Cycling", as well as two topics ("Rugby" and "Boxing") with a probability of 0.263 and 0.249 respectively. We also observed a "Cycling doping" topic which had a low probability (0.035).

The 4x500 condition: The four topics extracted ("Football", "Rugby", "Cycling", "Boxing") coincided with the nature of

[2] Examples originally conducted in French are presented in English for better comprehension.
[3]Similar results were found with 16, 24 and 32 words but we have limited our scope to the 12 most probable words.

the four sub-corpora and the probabilities were close to 0.250.

In the maximum similarity condition, the Topic model thus requires a minimum size of sub-corpora (number of documents) to extract relevant themes

Conclusion

In response to question 1, the results obtained in both 1a and 1b tests allow us to conclude that:

- the Topic model allows us to accurately extract topics congruent to the content of constituent sub-corpora and with a probability in line with the weight of these sub-corpora.
- the semantic similarity across the sub-corpora weakens the capacity of the Topic model to extract differentiated topics.
- increasing the number of documents from the same sphere of activity ("Sport") improves the capacity for extraction of relevant topics.

B. Test 2: Influence of the relative weight of sub-corpora on extracted topics.

In the previous tests, the corpora used were consistently composed of sub-corpora comprising an equal number of documents (40, 250 or 500). In the maximum similarity condition, extracted topics were relevant if sub-corpora were composed of 500 documents.

With the maximum similarity condition in mind, we sought to determine the extent to which varying the relative weight of the four sub-corpora affected the four extracted topics. If the relative weight indeed affected the extracted topics, then the topics would accurately reflect the nature of sub-corpora and their probability would vary according to their relative weight.

Material

From the 4x500 "Sports" similarity corpus previously used, two variants, "Sport-SP1" and "Sport-SP2", comprising an equal total number of documents (n=1190) but composed of four inversely proportional sub-corpora were constructed (Table 5).

Table 5: Number of documents comprising the Sports-SP1 and Sports-SP2 corpus

Corpus	Boxing	Cycling	Football	Rugby	Total
Sport SP1	500	360	230	100	**1190**
Sport SP2	100	230	360	500	**1190**

Results

The topics extracted in the "Sports-SP1" and "Sports-SP2" conditions were similar and reflected the nature of sub-corpora: "Boxing", "Cycling", "Football" and "Rugby" (see Table 5). Their probability was equal to the relative weight of sub-corpora (Figure 1). There were intrusions in the least probable topics extracted from the Sports-SP1corpus.

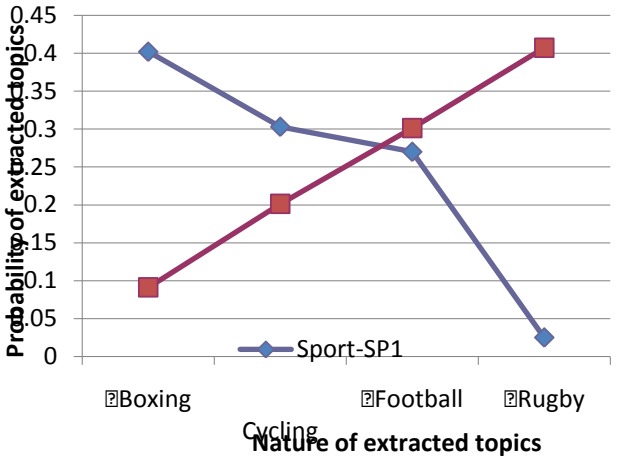

Figure 1: Nature and probability of extracted topics for the SP1 and SP2 "Sport" corpus.

Conclusion

In sum, Test 2 results show that, when sub-corpora are in variable proportions,

- the Topic model can accurately extract topics from a given relative weight of sub-corpora in the test corpus.
- the probability of extracted topics varies depending on the relative weight of sub-corpora.

C. Test 3: Permeability to the nature and scope of the context.

Test 3a.Permeability to the nature of the context

The previous tests have shown that the extraction of relevant topics is influenced by:

- similarity, which can be corrected by the provision of a minimum number of documents (Test 1),
- the relative weight of sub-corpora (Test 2).

Does the nature of surrounding corpora influence the quality of extracted topics? To respond to this question, we retained the "Sports" corpus (4x500 documents) previously used and formed four sub-corpora "Football", "Rugby", "Boxing" and "Cycling" before adding three corpora of a different nature but similar size (2000 documents). Six topics were extracted in order to determine the extent to which – when kept constant – the sub-corpora constituting the "Sports" corpus were still represented in extracted topics.

Material

Three corpora comprising 2000 documents were constructed from excerpts of:

- the "*Le Monde* 1999" newspaper, or
- the Encyclopedia (Biology), or
- "Literature" corpus[4] (documents with sad or happy connotations).

The characteristics of these three corpora are presented in Table 7.

[4] See the"French-Literature" semantic space on the lsa.colorado.edu website(Denhière, Lemaire, Bellissens&Jhean-Larose, 2007).

Three corpora were built around the core-corpus "Sports":
- 4x500 "Sport" Corpus + "*Le Monde* 99" Corpus
- 4x500 "Sport" Corpus + "Encyclopedia" Corpus
- 4x500 "Sport" Corpus + "Literature" Corpus

Table 7: Characteristics of the "*Le Monde* 1999", "Encyclopedia" and "Literature" corpora.

Corpus	Documents	Words	Different words
"*Le Monde* 99"	**2000**	**107 721**	**26 739**
"Encyclopedia"	**2000**	**105 222**	**22 437**
"Literature"	**2000**	**109 271**	**23 214**

Results

The results presented in Table 5 show that:
- the number of topics related to sports remained stable (three or four)
- while in two conditions ("Sports & Encyclopedia" and "Sports & Literature"), four extracted topics corresponded to four "Sports" sub-corpora ("Boxing", "Football", "Cycling" and "Rugby"), andin one condition ("Sports" & "*Le Monde* 99"), "Rugby" and "Football" were grouped under the same topic,
- the sum of probabilities of "Sports" topics barely varied from the nature of the added context (0.513, 0.517 and 0.510 respectively).

In terms of both quality and quantity,adding a corpus of a different nature barely modified the nature of the "Sports" topics extracted.

Conclusion

Whatever the nature of the corpus added to the core-corpus kept constant, the Topic model was able to extract topics reflecting fixed sub-corpora whose cumulative probability (0.500) was close to their weight in the corpus (50%).

The nature of neighboring corpora thus has little impact on the extraction of topics relating to a given corpus.

D. Test 3b: Permeability to the scope of the context.

Test 3a showed that the extraction of topics relating to a given corpus was barely influenced by neighboring corpora whose size was kept constant. But what would happen were the size of the corpora surrounding a given corpus to be progressively increased.

To respond to this question, the three corpora previously used ("*Le Monde* 99", "Encyclopedia" and "Literature") were successively added to the "Sports" corpus and thus represented respectively 1/2, 1/3 and 1/4 of the total corpus.

As with the previous test, six themes were extracted in order to determine the extent to which the sub-corpora constituting the "Sports" corpus, kept constant, continued to be represented in extracted topics.

1) Material

Three conditions corresponding to increasing corpora sizes were established:

Condition 1 (4000 documents):
"Sports" + "*Le Monde* 99 corpus"

Condition 2 (6000 documents):
"Sports" + "*Le Monde* 99" + "Encyclopedia" corpus

Condition 3 (8000 documents):
"Sports" + "*Le Monde* 99" + "Encyclopedia" + "Literature" corpus

Results

Table 9 shows the results of the extraction of six topics in the three experimental conditions.

In **condition one** (see above), three topics focused exclusively on sports – "Rugby and Football" (0.256), "Boxing" (0.151), "Cycling" (0.103) – and their cumulative probability accurately reflected their weight in the corpus.

In **condition two**, two topics focused exclusively on sports – "Football & Rugby" (0.170) and "Boxing & Cycling" (0.163) – and their cumulative probability (0.333) corresponded to their weight in the corpus.

In **condition three**, topic 2 – "Football, Boxing & Cycling"–focused exclusively on sports (0.164) and, as in the other conditions, its probability reflected its weight in the corpus. The term "Rugby" was found in the "Policy" topic (0.349).

In the three conditions, the extracted "Sports" topics thus accurately reflected the nature of sub-corpora and their relative weight within the total corpus

Conclusion

The extraction of topics related to a given corpus is barely influenced by the size of the corpus in which it is included (proportions of 1/2, 1/3 and 1/4). Except for condition three in which the term "Rugby" appeared in the "Policy" topic, the Topic model was capable of extracting topics reflecting fixed sub-corpora whose cumulative probability was close to their weight in the corpus (0.500, 0.333 and 0.250).

In conclusion, the themes extracted by the Topic model are little permeable to the nature and size of contexts added to the corpus.

IV. Conclusion

We carried out a comprehensive test on the capacity of the Topic model to extract topics in different corpus configurations. For each distribution, we presented the 12 words most likely to appear in an extracted topic. The first series of tests sought to evaluate how the size of a corpus influenced the topics extracted. Our results showed that the Topic model is capable of extracting topics from corpora and is hardly influenced by the density of the topic introduced in the overall corpus.

The objective of the second series of tests was to include a reference corpus within different corpora and observe how

the propagation of topics in the corpus influenced the extraction of topics using the Topic model. Results showed that the model (i) was able to extract homogenous topics regardless of their density in the overall corpus, and (ii) grouped topics with respect to a reference distribution reflecting the model's ability to disregard concepts introduced into the corpus.

We have therefore presented a semantic representation model that allows us to account for inferential mechanisms around semantic categories known as "topics" (Steyvers& Griffiths, 2007). We have demonstrated the model's capacity for abstraction as well as its ability to highlight homogenous conceptual groups from controlled reference corpora.

This model of semantic representation might be an alternative to Latent Semantic Analysis (Landauer&Dumais, 1997) in the conceptual instrumentation of text comprehension in reference to the Kintsch's (1998) construction-integration model. Specifically, it offers the possibility of identifying meaningful latent variables (topics) and also natively implements the asymmetry of associative memory, providing new perspectives for modeling syntactic rules (Kintsch&Mangalath, 2011).

Future research must clarify the relevance of the model beyond semantic categorization, i.e., beyond the level of topic construction, and explore how the Topic model makes it possible to account for an orientated organization of memory concepts. Indeed, while one of the limitations of LSA lies in the symmetry of association because of the use of the cosine to account for similarity, this limitation is overcome by the Topic model through the use of conditional probability. By extension, it thus becomes possible to natively integrate, using this model of semantic representation, the order of words and the functions of memory and forgetfulness within models for understanding.

Table 2: First words (n=12) from the four topics extracted for the three modalities of corpus similarity (4x40). It is worth mentioning that the authors are responsible for naming the topics.

Non-similarity corpus, ("Iron age, Breathing, Video games, Cycling")

Topic 1: "*Breathing*" $P(z_1) = .260$		Topic 2: "*Iron age*" $P(z_2) = .251$		Topic 3: "*Video games*" $P(z_3) = .248$		Topic 4: "*Cycling*" $P(z_4) = .241$	
Breathing	.010	**Copper**	.009	**Nintendo**	.009	**Cycling**	.008
Gills	.004	**Metal**	.008	Console	.005	Tour_de_France	.004
Oxygen	.004	Iron	.008	Ds	.005	Alberto_Contador	.003
Lungs	.004	Age	.006	Dsi	.004	Yellow_jersey	.002
Respiratory	.003	Millennium	.005	Sony	.004	Racer	.002
Environment	.003	Objects	.005	Wii	.004	Saxo_Bank	.002
Respiratory_tract	.003	Bronze	.004	Games	.003	Lance_Armstrong	.002
Muscles	.003	Metallurgy	.004	Ps3	.002	Rinaldo_Nocentini	.002
Air	.003	Jesus-Christ	.004	Playstation	.002	Arrival	.002
Animals	.002	Years	.003	Sales	.002	Team	.002
Divers	.002	Techniques	.002	Consoles	.002	Contador	.002
Function	.002	Metal	.002	Ipod_touch	.002	Won	.001
Water	.002	Tin	.002	Release	.001	Armstrong	.001

2*2 similarity corpus ("Breathing-Digestion, Football-Rugby")

Topic 1: "*Rugby*" $P(z_1) = .274$		Topic 2: "*Digestion*" $P(z_2) = .270$		Topic 3: "*Breathing*" $P(z_3) = .233$		Topic 4: "*Football*" $P(z_4) = .223$	
Rugby	.006	**Digestion**	.011	**Breathing**	.011	**Football**	.006
England	.003	Digestive_system	.007	Oxygen	.005	World_cup	.004
Top-14	.003	Foods	.006	Gills	.005	France	.003
Clermont	.003	Water	.005	Lung	.004	Ireland	.002
Face	.002	Stomach	.005	Respiratory	.003	FIFA	.002
Football	.002	Cells	.005	Respiratory_tracts	.003	Rugby	.002
Saturday	.002	Digestive-tract	.004	Muscles	.003	ACN_2010	.002
Sunday	.002	Enzymes	.003	Environment	.003	Angola	.002
Heineken-cup	.002	Small_intestine	.003	Air	.003	Match	.002
Victory	.002	Role	.003	Divers	.002	New-Zealand	.002
Club	.002	Blood	.003	Animals	.002	Center	.001
Scotland	.002	Circulation	.003	Function	.002	Striker	.001
Team	.002	Gastric	.003	Water	.002	Spain	.001

Maximum similarity corpus("Boxing – Cycling – Football – Rugby")

Topic 1: "*Boxing + Football*" $P(z_1) = .282$		Topic 2: "*Cycling + Rugby + Football*" $P(z_2) = .267$		Topic 3: "*Rugby + Cycling + Football*" $P(z_3) = .256$		Topic 4: "*Rugby + Cycling + Football*" $P(z_4) = .194$	
Boxing	.008	Cycling	.004	Rugby	.004	Rugby	.003
Face	.004	Tour_de_France	.003	Saturday	.003	Cycling	.002
Football	.004	Rugby	.003	Departure	.003	Saturday	.002
World_cup	.004	Victory	.002	Cycling	.002	Cardiff	.002
Fight	.003	Football	.002	Match	.002	Team	.002
Referee_stoppage	.003	Sunday	.002	Football	.002	Series	.002
American	.002	FIFA	.002	XV	.002	Rest	.001
Saturday	.002	Face	.002	Team	.002	Australia	.001
Dominant	.002	Italian	.002	French_Rugby_team	.002	Choice	.001
Dominated	.002	Wednesday	.002	Samoa	.002	Announced	.001
Scoring	.002	Conferred	.002	Weigh-in	.002	Positive	.001
ACN_2010	.002	Match	.002	Announced	.002	Football	.001
Friendly	.002	Encounter	.001	Encounter	.002	Player	.001

Table 3: First words (n=12) from the four topics extracted across the three modalities of corpus similarity 4x40, 4x250 and 4x500 from the Sport corpus, maximum similarity. It is worth mentioning that the authors are responsible for naming the topics.

Maximum similarity corpus 4x40

Topic 1: ***"Boxing + Football"*** $P(z_1) = .282$		**Topic 2:** ***"Cycling +Rugby + Football"*** $P(z_2) = .267$		**Topic 3:** ***"Rugby + Cycling + Football"*** $P(z_3) = .256$		**Topic 4:** ***"Rugby+ Cycling +Football"*** $P(z_4) = .194$	
Boxing	*.008*	Cycling	*.004*	Rugby	*.004*	Rugby	*.003*
Face	*.004*	Tour_de_France	*.003*	Saturday	*.003*	Cycling	*.002*
Football	*.004*	Rugby	*.003*	Departure	*.003*	Saturday	*.002*
World_cup	*.004*	Victory	*.002*	Cycling	*.002*	Cardiff	*.002*
Fight	*.003*	Football	*.002*	Match	*.002*	Team	*.002*
Referee_stoppage	*.003*	Sunday	*.002*	Football	*.002*	Series	*.002*
American	*.002*	FIFA	*.002*	XV	*.002*	Rest	*.001*
Saturday	*.002*	Face	*.002*	Team	*.002*	Australia	*.001*
Dominant	*.002*	Italian	*.002*	French_Rugby_team	*.002*	Choice	*.001*
Dominated	*.002*	Wednesday	*.002*	Samoa	*.002*	Announcement	*.001*
Scoring	*.002*	Conferred	*.002*	Weigh_in	*.002*	Positive	*.001*
ACN_2010	*.002*	Match	*.002*	Announced	*.002*	Football	*.001*
Friendly	*.002*	Encounter	*.001*	Encounter	*.002*	Player	*.001*

Maximum similarity corpus 4x250

Topic 1: ***"Football + Cycling"*** $P(z_1) = .452$		**Topic 2:** ***"Rugby"*** $P(z_2) = .263$		**Topic 3:** ***"Boxing"*** $P(z_3) = .249$		**Topic 4:** ***"Doping, Cycling"*** $P(z_4) = .035$	
Football	*.009*	Rugby	*.012*	Boxing	*.012*	Former	*.001*
Cycling	*.009*	Face	*.004*	Fight	*.008*	Doping	*.001*
World_cup	*.006*	Top-14	*.004*	Scoring	*.003*	Cycling	*.001*
Tour_de_France	*.005*	Saturday	*.003*	Face	*.003*	Champion	*.001*
Wednesday	*.003*	Match	*.003*	American	*.003*	Cyclist	*.001*
Transfers	*.003*	6_nations	*.003*	French	*.003*	Ama	*.000*
Team	*.002*	Ireland	*.002*	Heavy-weight	*.002*	Last	*.000*
Won	*.002*	French_Rugby_team	*.002*	Mexican	*.002*	Winner	*.000*
2010_world_cup	*.002*	France	*.002*	Featherweight	*.002*	June	*.000*
Saturday	*.002*	Heineken-cup	*.002*	Saturday_evening	*.002*	Friday	*.000*
Sunday	*.002*	Players	*.002*	Saturday	*.002*	Tested_positive	*.000*
Announced	*.002*	French_stadium	*.002*	Opponent	*.002*	Dimitri	*.000*

Maximum similarity corpus 4x500

Topic 1: ***"Cycling"*** $P(z_1) = .252$		**Topic 2:** ***"Rugby"*** $P(z_2) = .251$		**Topic 3:** ***"Football"*** $P(z_3) = .250$		**Topic 4:** ***"Boxing"*** $P(z_4) = .247$	
Cycling	*.014*	Rugby	*.014*	Football	*.015*	Boxing	*.014*
Tour_de_France	*.008*	Saturday	*.005*	World_cup	*.008*	Fight	*.008*
Won	*.003*	Face	*.004*	Transfers	*.005*	Scoring	*.004*
Team	*.003*	Top-14	*.004*	2010_world_cup	*.003*	Face	*.004*
Lance_Armstrong	*.003*	Match	*.003*	Wednesday	*.002*	American	*.003*
Racer	*.002*	6_nations	*.003*	Striker	*.002*	Heavy-weight	*.003*
Vuelta	*.002*	Heineken-cup	*.003*	Club	*.002*	Was_won	*.002*
Alberto_Contador	*.002*	French_Rugby_team	*.003*	Player	*.002*	Saturday_evening	*.002*
Arrival	*.002*	Ireland	*.002*	Face	*.002*	French	*.002*
Tour	*.002*	Players	*.002*	FIFA	*.002*	Opponent	*.002*
French	*.002*	French_stadium	*.002*	France	*.002*	Mexican	*.002*
Departure	*.002*	Team	*.002*	Midfield_player	*.002*	Saturday	*.002*
Astana	*.002*	Italy	*.002*	Saturday	*.002*	Referee_stoppage	*.002*

Table 6: First words (n=12) from the four topics extracted from the "Sports SP1" and "Sport SP2" corpora. It is worth mentioning that the authors are responsible for naming the topics.

Sport-SP1 Corpus: "Boxing (500), Cycling (360), Football (230), Rugby (100)"

Topic 1: "Boxing" $P(z_1) = .402$		**Topic 2: "Cycling"** $P(z_2) = .303$		**Topic 3: "Football"** $P(z_3) = .270$		**Topic 4: "Rugby"** $P(z_4) = .025$	
Boxing	*.017*	**Cycling**	*.014*	**Football**	*.010*	**FFR**	*.000*
Fight	*.010*	Tour_de_France	*.008*	World_cup	*.006*	Urban_boxing_united	*.000*
Face	*.004*	Won	*.003*	Rugby	*.004*	LNR	*.000*
Scoring	*.004*	Team	*.003*	Transfers	*.003*	Internet	*.000*
American	*.004*	Racer	*.003*	Face	*.003*	Departure	*.000*
Heavy-weight	*.003*	Lance_Armstrong	*.002*	Saturday	*.002*	Former	*.000*
Was_won	*.003*	Vuelta	*.002*	Match	*.002*	Press	*.000*
French	*.003*	Alberto_Contador	*.002*	Player	*.002*	National_league_of_Rugby	*.000*
Saturday_evening	*.002*	Tour	*.002*	Club	*.002*	Camp	*.000*
Mexican	*.002*	Departure	*.002*	France	*.002*	Mathis	*.000*
Opponent	*.002*	Winner	*.002*	Wednesday	*.002*	NetBoxing	*.000*
Referee_stoppage	*.002*	Friday	*.002*	2010_world_cup	*.002*	Meeting	*.000*

Sport-SP2 Corpus: "Boxing (100), Cycling (230), Football (360), Rugby (500)"

Topic 1: "Rugby" $P(z_1) = .407$		**Topic 2: "Football"** $P(z_2) = .301$		**Topic 3: "Cycling"** $P(z_3) = .201$		**Topic 4: "Boxing"** $P(z_4) = .091$	
Rugby	*.016*	Football	*.014*	Cycling	*.011*	Boxing	*.006*
Saturday	*.005*	World_cup	*.009*	Tour_de_France	*.005*	Fight	*.004*
Face	*.005*	Transfers	*.004*	Won	*.003*	American	*.002*
Top-14	*.005*	2010_world_cup	*.003*	Team	*.002*	Scoring	*.002*
Match	*.004*	Wednesday	*.003*	Lance_Armstrong	*.002*	Heavy-weight	*.001*
6_nations	*.003*	Striker	*.002*	Alberto_Contador	*.002*	Face	*.001*
Heineken-cup	*.003*	Club	*.002*	Racer	*.002*	Was_won	*.001*
French_Rugby_team	*.003*	FIFA	*.002*	Vuelta	*.002*	Saturday_evening	*.001*
Italy	*.002*	France	*.002*	French	*.002*	Venue	*.001*
French_stadium	*.002*	Player	*.002*	Winner	*.002*	Meeting	*.001*
Team	*.002*	Face	*.002*	Arrival	*.001*	Dominant	*.001*
Players	*.002*	Players	*.002*	Wednesday	*.001*	Undefeated	*.001*

Table 8: First words from the topics resulting from the "Sports &*Le Monde* 99", "Sports & Encyclopedia" and "Sports & Literature" corpora based on a distribution across six topics.It is worth mentioning that the authors are responsible for naming the topics.

"Sports &*Le Monde* 99" Corpus

Topic 1: "French policy" $P(z_1) = .403$		**Topic 2: "Rugby and Football"** $P(z_2) = .256$		**Topic 3: "Boxing"** $P(z_3) = .151$		**Topic 4: "Cycling"** $P(z_4) = .103$		**Topic 5: "Arts"** $P(z_5) = .068$		**Topic 6: "Spectacles"** $P(z_5) = .019$	
France	.002	Rugby	.007	Boxing	.009	Cycling	.008	History	.001	Tel	.001
Country	.002	Football	.007	Fight	.005	Tour_de_France	.005	Century	.001	Theater	.000
President	.002	World_cup	.004	American	.002	Lance_Armstrong	.002	Film	.001	Children	.000
Policy	.002	Saturday	.003	Face	.002	Team	.001	Scene	.001	Woods	.000
World	.002	Face	.003	Scoring	.002	Vuelta	.001	Art	.001	Road	.000
Government	.002	Match	.002	French	.002	Alberto_Contador	.001	Death	.001	Foot	.000
French	.002	Transfers	.002	Heavy-weight	.002	Arrival	.001	Man	.001	Poster	.000
Paris	.001	Top-14	.002	Saturday	.001	Won	.001	City	.000	Venues	.000
State	.001	Club	.002	Cycling	.001	Racer	.001	Roman	.000	PS	.000
Today	.001	Players	.002	Was_won	.001	Tour	.001	Years	.000	Stations	.000
Minister	.001	Wednesday	.002	Saturday_evening	.001	Departure	.001	Music	.000	Opera	.000
Former	.001	Player	.002	Opponent	.001	Racers	.001	Woman	.000	Carry_away	.000
Work	.001	Team	.002	Mexican	.001	Race	.001	Father	.000	Propose	.000

"Sports & Encyclopedia" Corpus

Topic 1: "Breathing" $P(z_1) = .322$		**Topic 2: "Circulation"** $P(z_2) = .161$		**Topic 3: "Football"** $P(z_3) = .134$		**Topic 4: "Rugby"** $P(z_4) = .130$		**Topic 5: "Boxing"** $P(z_5) = .129$		**Topic 6: "Cycling"** $P(z_5) = .124$	
Water	.005	Blood	.002	Football	.010	Rugby	.010	Boxing	.010	Tour	.014
Air	.003	Years	.002	Cup	.009	France	.005	Fight	.008	Cycling	.010
Blood	.003	Heart	.002	World	.009	Saturday	.005	Champion	.006	France	.009
Oxygen	.003	King	.002	Transfers	.003	XV	.005	Title	.006	Phase	.007
Body	.002	Days	.001	Match	.003	Match	.004	KO	.005	Team	.004
Earth	.002	Day	.001	Club	.003	Face	.004	World	.004	Racer	.003
Man	.002	Name	.001	Wednesday	.002	Top-14	.003	Weights	.004	Armstrong	.003
Form	.002	French	.001	World	.002	Stadium	.003	Years	.004	Won	.002
Breathing	.002	Death	.001	Striker	.002	Nations	.003	Belt	.003	Training	.002
Energy	.002	English	.001	Season	.002	VI	.003	WBA	.003	Years	.002
Organism	.002	Mans	.001	Players	.002	Team	.002	Points	.003	Jersey	.002
Family	.002	France	.001	ACN	.002	Players	.002	Heavy-weight	.003	Contador	.002
Meters	.002	Big	.001	Selection	.002	Group	.002	Saturday	.003	Season	.002

"Sports & Literature" Corpus

Topic 1: "Emotions" $P(z_1) = .432$		**Topic 2: "Football"** $P(z_2) = .166$		**Topic 3: "Boxing"** $P(z_3) = .128$		**Topic 4: "Rugby"** $P(z_4) = .128$		**Topic 5: "Cycling"** $P(z_5) = .091$		**Topic 6: "Emotions"** $P(z_5) = .055$	
Sorrow	.004	World	.009	Boxing	.010	Rugby	.010	Tour	.014	Joy	.001
Joy	.004	Football	.009	Fight	.007	Saturday	.005	France	.008	Little	.001
Man	.004	Cup	.008	Champion	.006	France	.005	Cycling	.008	Sad	.001
Life	.003	Season	.003	Title	.005	XV	.005	Phase	.007	Chick	.001
Time	.003	Years	.003	KO	.005	Match	.004	Armstrong	.003	Sorrow	.001
Sad	.003	Team	.003	World	.004	Face	.004	Won	.002	Merry	.001
Day	.003	Transfers	.003	Weights	.004	Top-14	.003	Team	.002	Woods	.001
Sir	.002	Match	.002	Years	.003	Nations	.003	Jersey	.002	Jo	.001
Eyes	.002	Club	.002	Belt	.003	Stadium	.003	Contador	.002	Doctor	.000
Madam	.002	Contract	.002	WBA	.003	VI	.003	Vuelta	.002	Fire	.000
Times	.002	France	.002	Points	.003	Players	.002	Lance	.002	Cup	.000
Big	.002	Wednesday	.002	Heavy-weight	.003	Team	.002	Racer	.002	Eyes	.000
Woman	.002	Cycling	.002	Saturday	.003	Group	.002	Training	.002	Sky	.000

Table 9:First words from the topics resulting from the "Sports &*Le Monde* 99" "Sports, *Le Monde* 99 & Encyclopedia" and "Sports, *Le Monde* 99, Encyclopedia & Literature" corpora based on a distribution across six topics. It is worth mentioning that the authors are responsible for naming the topics.

"Sports &*Le Monde* 99" corpus

Topic 1: "Policy France" $P(z_1) = .403$		**Topic 2: "Rugby & Football"** $P(z_2) = .256$		**Topic 3: "Boxing"** $P(z_3) = .151$		**Topic 4: "Cycling"** $P(z_4) = .103$		**Topic 5: "Arts"** $P(z_5) = .068$		**Topic 6: "Spectacles"** $P(z_5) = .019$	
France	.002	Rugby	.007	Boxing	.009	Cycling	.008	History	.001	Tel	.001
Country	.002	Football	.007	Fight	.005	Tour_de_France	.005	Century	.001	Theater	.000
President	.002	World_cup	.004	American	.002	Lance_Armstrong	.002	Film	.001	Children	.000
Policy	.002	Saturday	.003	Face	.002	Team	.001	Scene	.001	Woods	.000
World	.002	Face	.003	Scoring	.002	Vuelta	.001	Art	.001	Road	.000
Government	.002	Match	.002	French	.002	Alberto_Contador	.001	Death	.001	Foot	.000
French	.002	Transfers	.002	Heavy-weight	.002	Arrival	.001	Man	.001	Poster	.000
Paris	.001	Top-14	.002	Saturday	.001	Won	.001	City	.000	Venues	.000
State	.001	Club	.002	Cycling	.001	Racer	.001	Roman	.000	PS	.000
Today	.001	Players	.002	Was_won	.001	Tour	.001	Years	.000	Stations	.000
Minister	.001	Wednesday	.002	Saturday_evening	.001	Departure	.001	Music	.000	Opera	.000
Former	.001	Player	.002	Opponent	.001	Racers	.001	Woman	.000	Carry_away	.000
Work	.001	Team	.002	Mexican	.001	Race	.001	Father	.000	Propose	.000

"Sports, *Le Monde* 99 &Encyclopedia" corpus

Topic 1 "Breathing" $P(z_1) = .263$		**Topic 2: "Policy"** $P(z_2) = .261$		**Topic 3 "Football & Rugby"** $P(z_3) = .170$		**Topic 4 "Boxing & Cycling"** $P(z_4) = .163$		**Topic 5 "Policy"** $P(z_5) = .133$		**Topic 6 "Press"** $P(z_5) = .010$	
Water	.003	France	.002	Rugby	.006	Boxing	.006	King	.001	Tel	.000
Blood	.003	President	.001	Football	.005	Cycling	.005	Day	.001	Correct	.000
Air	.002	Country	.001	World_cup	.003	Fight	.003	City	.001	Numbers	.000
Body	.001	Policy	.001	Face	.002	Tour_de_France	.003	Big	.001	Complementary	.000
Organism	.001	French	.001	Saturday	.002	American	.001	Death	.001	Thrush	.000
Oxygen	.001	Government	.001	Match	.002	French	.001	World	.001	Shepherds	.000
Man	.001	World	.001	Transfers	.001	Face	.001	Name	.001	Goats	.000
Earth	.001	State	.001	Top-14	.001	Won	.001	France	.001	Melodies	.000
Breathing	.001	Paris	.001	Club	.001	Scoring	.001	Man	.000	Arkestra	.000
Name	.001	Today	.001	Players	.001	Saturday	.001	War	.000	Road	.000
Family	.001	Minister	.001	Player	.001	Heavy-weight	.001	Paris	.000	Children	.000
Cells	.001	Former	.000	Team	.001	Victory	.001	French	.000	Sold	.000

"Sports, *Le Monde* 99, Encyclopedia & Sad Literature"SP8000 corpus

Topic 1 "Policy" $P(z_1) = .349$		**Topic 2 "Emotions"** $P(z_2) = .256$		**Topic 3 "Breathing"** $P(z_3) = .180$		**Topic 4 "Football, Boxing & Cycling"** $P(z_4) = .164$		**Topic 5 "Emotions"** $P(z_5) = .032$		**Topic 6 "Arts"** $P(z_6) = .019$	
Rugby	.002	Sorrow	.002	Water	.003	Football	.004	Joy	.000	Min	.000
France	.002	Life	.002	Blood	.002	Boxing	.004	Woods	.000	Div	.000
Country	.001	Man	.002	Air	.002	Cycling	.004	Merry	.000	CD	.000
President	.001	Joy	.002	Body	.001	Fight	.002	Egyptians	.000	Fra	.000
Policy	.001	Day	.002	Organism	.001	World_cup	.002	Wind	.000	John	.000
Government	.001	Sad	.001	Oxygen	.001	Tour_de_France	.002	Sky	.000	Andrew	.000
French	.001	Death	.001	Earth	.001	Face	.001	Stars	.000	Hubert	.000
State	.001	Sir	.001	Name	.001	Saturday	.001	Flowers	.000	DVD	.000
Francs	.001	Eyes	.001	Man	.001	American	.001	Africa	.000	Music	.000
World	.001	World	.001	Breathing	.001	Transfers	.001	Spring	.000	Eastwood	.000
Today	.001	Time	.001	Energy	.001	French	.001	Wolves	.000	Gautier	.000
Paris	.001	Big	.001	Sea	.001	Victory	.001	Song	.000	Montreuil	.000

TECHNICAL APPENDIX

PRESENTATION OF THE TOPIC MODEL

In the topic model, each topic provides a probability distribution over a set of words, and each document provides a probability distribution over a set of topics. For a given set of *T* topics, the probability of the *i*thword of the document is expressed as:

$$P(w_i) = \sum_{j=0}^{T} P(w_i|z_i = j)P(z_i = j) \quad \text{[A-1.1]}$$

Where z_i is a latent variable which indicates the topic from which the word w_i is derived.

Within a document,
$P(w_i|z_i = j)$ is the probability of word w_i in topic j ;
$P(z_i = j)$ is the probability of assigning a word to topic j.

For *D* documents comprising *T* topics expressed using *n* unique words,

- $P(w \mid z)P(w|z)$ is denoted by a set of *T* multinomial distributions, $P(w|z_i = j) = \Phi_w^{(j)}$ $P(w|z = j) = \Phi_w^{(j)}$,
- $P(z)$by a set of *D* multinomial distributions, $P(z = j) = \theta_j^{(d)} P(z = j) = \theta_j^{(d)}$.

The identification of the topics used in a corpus with *n* unique words $\{w_1 \ldots, w_n\}$ can be obtained by estimating the Φ and Θ matrices:

Φdenotes the most important words in a given topic,
and Θ denotes the most important topics in a document within the corpus.

$\Phi$and Θ matrices are generated from the words ×documents co-occurrence matrix using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). The Φ x Θ matrix product provides the probability of occurrence of a word in a given document. The probability distribution over a set of words for each document in the corpus, *P(w | d)*, is estimated by the matrix product of probability distributions of topics over a set of words, *P(w | z)*, and by the probability distribution of documents over a set of topics, *P(z | d)* (Figure 1).

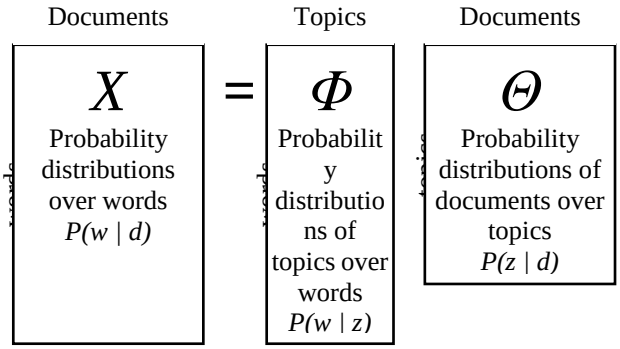

Figure 2: The decomposition of the words × documents probability distribution using Latent Dirichlet Allocation(Blei, et al., 2003)

In the light of these two distributions, the conditional probability of word w_1 given word w_2 can be obtained via:

$$P(w_1|w_2) = \frac{1}{P(w_2)} \sum_{j=0}^{T} P(w_1|z_j)P(w_2|z_j)P(z_j) \quad \text{[A-1.2]}$$

V. EXTRACTING TOPICS IN A SMALLER CORPUS

We will illustrate the three phases of topic extraction using the DEMO corpus that is composed of 12 documents (sentences). Six documents focus on the life of bees and the other six on the planets in the solar system (Table 10).

Table 10. Smaller DEMO corpus composed of six "bee" documents and six "planet" documents. The words that were used to define the co-occurrence matrix (at least two occurrences) are highlighted.

–	Document 1: **Bees live** in colonies, in a **hive**.
–	Document 2: Each **bee** colony has its **queen** which is bigger than the other bees.
–	Document 3: The **queenbee**'s duty is to lay eggs.
–	Document 4: **Worker bees** work all the time.
–	Document 5: **Worker bees** do not lay eggs.
–	Document 6: Male **bees** do **live** in the **hive** throughout the year
–	Document 7: **Mars** is called the red **planet**.
–	Document 8: **Mars** is a cold **planet**.
–	Document 9: **Neptune** is an ice-covered **planet**.
–	Document 10: **Planet Neptune** has very **faint rings**.
–	Document 11: **Saturn** is a **planet** surrounded by thousands of **rings**.
–	Document 12: **Planet Saturn'srings** are very faint.

The three phases involved:
- Generating the words × documents co-occurrence matrix,
- Choosing the number of topics to extract,
- Extracting Φ and Θ probability distributions.

1. Generating the co-occurrence matrix

The DEMO corpus was transformed into a co-occurrence matrix *M* with words in the corpus placed in the horizontal rows and documents in the columns (Table 11). We retained the words appearing at least twice and deleted all stop words. The *M(i,j)* intersection is the number of times word w_i appeared in document d_j. For instance, the word "hive" appeared in documents one and six and the word "rings" in documents 10, 11 and 12.

Table 11: The *M* co-occurrence matrix from the DEMO corpus.

	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8	Document 9	Document 10	Document 11	Document 12
Word 1: "bees"	1	1	1	1	1	1	0	0	0	0	0	0
Word 2: "live"	1	0	0	0	0	1	0	0	0	0	0	0
Word 3: "hive"	1	0	0	0	0	1	0	0	0	0	0	0
Word 4: "queen"	0	1	1	0	0	0	0	0	0	0	0	0
Word 5: "workers"	0	0	0	1	1	0	0	0	0	0	0	0
Word 6: "Mars"	0	0	0	0	0	0	1	1	0	0	0	0
Word 7: "planet"	0	0	0	0	0	0	1	1	1	1	1	1
Word 8: "Neptune"	0	0	0	0	0	0	0	0	1	1	0	0
Word 9: "rings"	0	0	0	0	0	0	0	0	0	1	1	1
Word 10: "faint"	0	0	0	0	0	0	0	0	0	1	0	1
Word 11: "Saturn"	0	0	0	0	0	0	0	0	0	0	1	1

2. Choosing the number of topics and generating the Φ *and* Θ *matrices*

Matrix *M* was manipulated using Latent Dirichlet Allocation (LDA), whose parameter was set to allow the extraction of two topics in order to generate *Φ* and *Θ* matrices (Table 12 and Table 13). We based our calculations on the compiled library developed by Andrzejewski, Mulhern, Liblit, & Zhu (2007).

Table 12: *Φ* matrix resulting from the application of the Latent Dirichlet Allocation to the co-occurrence matrix of the DEMO corpus. The highest probabilities for each topic are in italics and are highlighted in yellow.

	Topic 1	**Topic 2**
Word 1: "bees"	*0.269*	0.037
Word 2: "live"	*0.115*	0.037
Word 3: "hive"	*0.115*	0.037
Word 4: "queen"	*0.115*	0.037
Word 5: "workers"	*0.115*	0.037
Word 6: "Mars"	0.038	*0.111*
Word 7: "planet"	0.038	*0.259*
Word 8: "Neptune"	0.038	*0.111*
Word 9: "rings"	0.038	*0.148*
Word 10: "faint"	0.077	*0.074*
Word 11: "Saturn"	0.038	*0.111*

Table 13: *Θ* matrix resulting from the application of the Latent Dirichlet Allocation to the co-occurrence matrix of the DEMO corpus. The highest probabilities for each topic are in italics and are highlighted in yellow.

Document	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Topic 1	.97	.96	.96	.96	.96	.97	.05	.05	.05	.26	.03	.02
Topic 2	.03	.05	.05	.05	.05	.03	.96	.96	.96	.74	.97	.98

In the *Φ* matrix, the words "bees", "live", "hive", "queen" and "workers" from the first six documents have the highest probability in topic 1(from 0.115 to 0.269), and the words "Mars", "planet", "Neptune", "rings", "Saturn" and "faint" from the six last documents have the highest probability in topic 2 (from 0.074 to 0.259).

In the *Θ* matrix, topic 1 has the highest probability in the first six documents of the corpus (from 0.955 to 0.969), and topic 2 the highest probability in the last six documents (from 0.738 to 0.976).

The *Φ* and *Θ*matrix product (Table 6) estimates the probability of occurrence of a word in each document.

Table 14: *Φ* and *Θ* matrix products resulting from the application of the Latent Dirichlet Allocation to the co-occurrence matrix of the DEMO corpus.

Document	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Bees	*.262*	*.259*	*.259*	*.259*	*.259*	*.262*	*.048*	*.048*	*.048*	*.098*	*.044*	*.043*
Live	*.113*	*.112*	*.112*	*.112*	*.112*	*.113*	*.041*	*.041*	*.041*	*.058*	*.039*	*.039*
Hive	*.113*	*.112*	*.112*	*.112*	*.112*	*.113*	*.041*	*.041*	*.041*	*.058*	*.039*	*.039*
Queen	*.113*	*.112*	*.112*	*.112*	*.112*	*.113*	*.041*	*.041*	*.041*	*.058*	*.039*	*.039*
Workers	*.113*	*.112*	*.112*	*.112*	*.112*	*.113*	*.041*	*.041*	*.041*	*.058*	*.039*	*.039*
Mars	*.041*	*.042*	*.042*	*.042*	*.042*	*.041*	*.108*	*.108*	*.108*	*.092*	*.109*	*.109*
Planet	*.045*	*.048*	*.048*	*.048*	*.048*	*.045*	*.249*	*.249*	*.249*	*.201*	*.252*	*.254*
Neptune	*.041*	*.042*	*.042*	*.042*	*.042*	*.041*	*.108*	*.108*	*.108*	*.092*	*.109*	*.109*
Rings	*.042*	*.043*	*.043*	*.043*	*.043*	*.042*	*.143*	*.143*	*.143*	*.119*	*.145*	*.146*
Faint	*.077*	*.077*	*.077*	*.077*	*.077*	*.077*	*.074*	*.074*	*.074*	*.075*	*.074*	*.074*
Saturn	*.041*	*.042*	*.042*	*.042*	*.042*	*.041*	*.108*	*.108*	*.108*	*.092*	*.109*	*.109*

3. CONDITIONAL PROBABILITY AND ASSOCIATIVE ASYMMETRY

Using the *Φ* and *Θ* matrices and the [A-1.2] formula, we can calculate the conditional probability between two words (Table 6).

Intra- and inter-topic conditional probability

We assumed that the probability of association between two words belonging to different topics would be lower than the probability of association between two words belonging to the same topic. The data presented in Table 6 support this hypothesis: the probability of the word "bees" conditioned on the word "hive" (0.216), and the word "planet" conditioned on the word "rings" (0.211) was higher than the probability of the word "planet" conditioned on the word "bees" (0.047) and the probability of the word "hive" conditioned on the word "rings" (0.054).

Table 15: Intra- and inter-topic conditional probability.

Word 1	Word 2	Inter-topic	Intra-topic
Bees	Hive		0.216
Planet	Rings		0.211
Planet	Bees	0.047	
Hive	Rings	0.054	

Associative asymmetry

With reference to the corpus, we had posited that the probability of the word "planet" conditioned on the words "Mars", "Saturn" or "Neptune" would be higher than the probabilities of the words "Mars", "Saturn" or "Neptune" conditioned on the word "planet". The data presented in Table 16 support this assumption.

Table 16: Asymmetry of conditional probability

Word 1	Word 2	P(Word 1 | Word2)
Planet	Mars	0.199
Mars	Planet	0.101
Planet	Saturn	0.199
Saturn	Planet	0.101
Planet	Neptune	0.199
Neptune	Planet	0.101

REFERENCES

[1] Andrzejewski, D., Mulhern, A., Liblit, B., & Zhu, X. (2007). Statistical Debugging using Latent Topic Models. In: Proceedings of the 18th European Conference on Machine Learning (ECML 2007), 6-17.

[2] Barclay, J.R., Bransford, J.D., Franks, J.J. McCarrell, N.S., &Nitsch, K.E. (1974).Comprehension and semantic flexibility. Journal of Verbal Learning and Verbal Behavior, 13, 471-481.

[3] Bellissens C., Thérouanne, P., &Denhière G., (2004). Deuxmodèlesvectoriels de la mémoiresémantique : description, théorie et perspective. Le Langage et L'Homme, 39 (2), 101-121.

[4] Blanc, N., &Brouillet, D. (2003). Mémoire et compréhension. Lire pour comprendre. Psycho Press, 2003.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[6] Burgess, C., & Lund, K. (1997). Modeling cerebral asymetries of semantic memory using high-dimensional semantic space. In M. Beeman& C. Chiarello (Eds.), Right Hemisphere Language comprehension: Perspectives from cognitive neuroscience (pp. 117-156). Hillsdale, N.J.: Erlbaum Press.

[7] Denhière, G. &Tapiero, I. (1996). La signification comme structure émergeante: de l'accès au lexique à la compréhension de textes. In. V. Rialle& D. Fisette (Eds.), Penserl'esprit. Des sciences de la cognition à unephilosophie cognitive (pp. 307-336). Grenoble: PUG.

[8] Denhière, G., Lemaire, B., Bellissens, C., &Jhean-Larose, S. (2004). Psychologie cognitive et compréhension de textes. In. S. Porhiel& D. Klingler (Eds.). Regards croisés sur l'unitétexte (pp. 74-95). Pleyben: Perspectives.

[9] Denhière, G., Lemaire, B., Bellissens, C., &Jhean-Larose, S. (2007). A Semantic Space for Modeling Children's Semantic Memory. In. T. K. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis (pp. 143-166). Magwah, New Jersey: Lawrence Erlbaum Associate.

[10] Denhière, G., Leveau, N., &Jhean-Larose, S. (2010). Mémoire, représentationsémantiqueprobabiliste et extraction automatique de thèmes. 52ème congrès de la SociétéFrançaise de Psychologie, Cognition Emotion et Société. Lille, 7-9 Septembre.

[11] Dennis, S. (2005). A Memory-Based Theory of Verbal Cognition. Cognitive Science, 29, 145-193.

[12] Diaz, J., Rifqi, M., Bouchon-Meunier, B., Jhean-Larose, S. &Denhière, G. (2008). Imperfect Answer in multiple choice questionnaires. Proceedings of the Third European Conference on Technology Enhanced Learning (EC-TEL 08), 144-154.

[13] Griffiths, T. L., &Steyvers, M. (2003). Prediction and semantic association. Advances in Neural Information Processing Systems, 15.

[14] Griffiths, T. L., &Steyvers, M. (2004). Finding Scientific Topics. Paper presented at the National Academy of Sciences.

[15] Griffiths, T. L., Steyvers, M., &Tenenbaum, J. B. (2007). Topics in Semantic Representation. Psychological Review, 114(2), 211-244.

[16] Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, Instruments and Computers, 16, 96-101.

[17] Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. Paper presented at the Uncertainty in Artificial Intelligence, UAI'99, Stockholm.

[18] Jhean-Larose, S., &Denhière, G. (Eds.) (2010). Learning, Memory and Latent Semantic Analysis, StudiaInformaticaUniversalis, whole vol. 8 (1). Paris: Hermann Éditeurs.

[19] Jhean-Larose, S., Leclercq, V., Diaz, J., Denhière, G., &Bouchon-Meunier, B. (2010). Knowledge evaluation based on LSA: MCQs and free answer questions. In. S. Jhean-Larose, & G. Denhière (Eds.). Learning, Memory and Latent Semantic Analysis (pp. 53-78), StudiaInformaticaUniversalis. Paris: Hermann Éditeurs.

[20] Kintsch, W. (1988). The role of knowledge in discourse comprehension construction-integration model. Psychological Review, 95, 163-182.

[21] Kintsch, W. (1998). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

[22] Kintsch, W. (2000) Metaphor comprehension: A computational theory. Psychonomic Bulletin & Review, 7, 257-266.

[23] Kintsch, W. (2001). Predication. Cognitive Science, 25, 173-202.

[24] Kintsch, W. (2010). Modeling Semantic Memory. In. S. Jhean-Larose, & G. Denhière (Eds.). Learning, Memory and Latent Semantic Analysis (pp. 1-21), StudiaInformaticaUniversalis. Paris: Hermann Éditeurs.

[25] Kintsch, W., &Mangalath, P. (2011). The construction of meaning. Topics in Cognitive Science. 3, 346-370.

[26] Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production, Psychological Review, 85(5), 363-394.

[27] Landauer, T. K., &Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.

[28] Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In. M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 1 (pp. 45-51). Cambridge: MIT Press.

[29] Landauer, T. K., McNamara, D. S., Dennis, S., &Kintsch, W. (2007). Handbook of Latent Semantic Analysis. Magwah, New Jersey: Lawrence Erlbaum Associates.

[30] Lemaire, B., Denhière, G., Bellissens, C., &Jhean-Larose, S. (2006). A model of computer program for simulating text comprehension. Behavior Research Methods, Instruments and Computers, 38 (4). 628-637.

[31] Lifchitz, A., Jhean-Larose, S. &Denhière, G. (2009). Effect of tuned parameters on a LSA multiple choice questions answering model. Behavior Research Methods, 41 (4), 1201-1209.

[32] Murdock, B.B., Jr., (1982). A theory for the storage and retrieval of items and associative information, Psychological Review, 89, 609-626.

[33] Rumelhart, D., McClelland, J., & The PDP Research Group (1986). Parallel Distributed Processing: Explorations in the microstructure of cognition. Cambridge: MIT Press.

[34] Steyvers, M., & Griffiths, T. L. (2007). Probabilistic Topic Models. In T. K. Landauer, D. S.McNamara, S. Dennis & W. Kintsch (Eds.), Latent Semantic Analysis A Road to Meaning (pp. 427-448). Magwah, New Jersey: Lawrence Erlbaum Associate.

[35] Tiberghien, G. (1997). La mémoireoubliée. Liège: Mardaga.