Topic Models applied to Neuroscientific Literature

Erick Cobos
École Polytechnique Fédérale de Lausanne
erick.cobos@epfl.ch

Jean-Cédric Chappelier
École Polytechnique Fédérale de Lausanne
jean-cedric.chappelier@epfl.ch

Renaud Richardet
École Polytechnique Fédérale de Lausanne
renaud.richardet@epfl.ch

June 06, 2014

Abstract

In this project we applied standard topic models (Latent Dirichlet Allocation) to neuroscientific literature extracted from the PubMed database; we have replicated the results obtained in [21] and used the MeSH information from the PubMed articles to compute a correlation measure between topics and MeSH descriptors, MeSH graph nodes and MeSH tree numbers with different degrees of success. Finally, the preprocessing of documents was enhanced to include the handling of multiwords producing good results.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We offer an small introduction to concepts used in this project.

1.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation [2], also referred as *Topic Models* or *Discrete PCA models* [7], is a generative model used in Natural Language Processing that explains the ocurrence of observed data (words) by hidden or *latent variables* (topics). A *topic* is a probability distribution over the vocabulary. Intuitively, we expect topics to describe a specific concept, thus having high probability for those words most closely related to that concept. A topic related to cars, for example, could have high probabilities for words: "tire", "motor", and "wheel" and low probability for other non-related words.
In the standard topic model, a document is represented as a *mixture* of topics, it is, a probability distribution over all topics. A topic model is a mathematical framework that allows examining a set of documents and discovering, based on the statistics of the words in each, a list of topics and each document's topic distribution.
A more intuitive introduction to topic models is given on [11].

1.1.1 LDA Framework

In LDA, topic distributions per document $p(z|d)$ and word distributions per topic $p(w|z)$ are assumed to have a Dirichlet prior distribution (denoted by $Dir(\boldsymbol{\alpha})$ and $Dir(\boldsymbol{\beta})$) and are learned using Bayesian inference. The usual notation in LDA [2] is the following:

1. K is the number of topics
2. V is the number of words in the vocabulary
3. M is the number of documents in the corpus

4. N_i is the number of words in document i
5. $\boldsymbol{\theta}_i$ is the topic distribution for document i, where $\theta_i^{(j)} = p(z_j|d_i)$. Equally, $\boldsymbol{\theta}$ is the topic distribution matrix for all documents (dimension MxK)
6. $\boldsymbol{\phi}_k$ is the word distribution for topic k, where $\phi_k^{(j)} = p(w_j|z_k)$. Equally, $\boldsymbol{\phi}$ is the word distribution for all topics (dimension KxV)
7. $\boldsymbol{\alpha}$ is the K-dimensional vector with the parameters for the prior Dirichlet of the topic distribution per document $\boldsymbol{\theta}_i$ for all i
8. $\boldsymbol{\beta}$ is the V-dimensional vector with the parameters for the prior Dirichlet of the word distribution per topic $\boldsymbol{\phi}_i$ for all i. Usually all components of $\boldsymbol{\beta}$ have the same value reducing the prior to $Dir(\beta)$
9. w_{ij} is the j^{th} word on document i
10. z_{ij} is the topic assigned by the model to w_{ij}

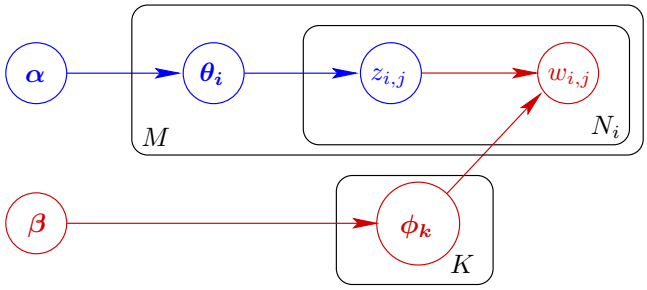


Figure 1.1: Graphical model representation of LDA. Boxes are "plates" representing replicates.

1.1.2 Software

For this project we use the *Discrete Component Analysis* (DCA) software [3] to train our LDA model. The program receives the corpus of preprocessed documents, allows you to specify the $\boldsymbol{\alpha}$ and β for the priors and the number of topics and uses a random sampling method to estimate the parameters for the topic model [4]. Parameters as the topic distribution per document $\boldsymbol{\theta}$ or the word distribution per topic $\boldsymbol{\phi}$ can later be retrieved from the output model. This program was chosen after a careful evaluation of different options for topic models that considered efficiency, functionality and scalability [21].
HCA [6], a research software by the author of DCA, was made available in 2013 and could be used for later stages of the project. HCA does various versions of non-parametric topic models including LDA, HDP-LDA and NP-LDA.

1.2 Medical Subject Headings (MeSH)

MeSH is a comprehensive medical vocabulary managed by the United States National Library of Medicine. It consists of descriptors arranged in a hierarchical structure and is primarily used for indexing journal articles and books in the life sciences. As any descriptor has a list of similar terms it can also serve as a thesaurus that facilitates searching. The MeSH vocabulary is continually revised and updated by the Medical Subject Headings Section staff from the Library of Medicine. We offer a short introduction to the MeSH structure and its use in the PubMed database.

1.2.1 Subject headings

Subject headings, *main headings* or *descriptors* are the principal components of the MeSH hierarchy. They are used to index articles from 5,400 of the world's leading biomedical journals for the PubMed database [13]. Every descriptor is accompanied by a short definition, links to related descriptors and a set of synonyms or *entry terms*. They are generally updated on an annual basis to reflect changes in vocabulary and additions to the medical literature. There are 27 149 descriptors in the 2014 MeSH vocabulary, each one of which has a Unique Identifier assigned to it (starting with D and followed by 6 to 9 digits).

1.2.2 Structure

Descriptors are arranged in a twelve-level hierarchy where the most general terms as "Body Regions" or "Mental Disorders" appear in higher levels and more specific headings are found at deeper levels. The first level is composed of sixteen *categories* represented by a capital letter: "Anatomy" [A], "Organisms" [B], "Diseases" [C], "Chemicals and Drugs" [D], etc. Categories serve as an initial division for MeSH descriptors but are not descriptors themselves.
Each MeSH descriptor appears in at least one place in the hierarchy (sometimes referred as a "tree") and may appear in as many additional places as may be appropriate. Every appearance in the hierarchy is uniquely represented by a *tree number*. Therefore, each MeSH descriptor can have various tree numbers.
The MeSH structure can be viewed as a directed acyclic graph with nodes representing a single MeSH descriptor and directed edges representing the parent-child relation between descriptors. In this case, each graph node consists of a MeSH descriptor and a set of tree numbers, one for each place in the hierarchy where the descriptor appears.
For example, the MeSH descriptor "Body Regions" with Unique ID D001829 has tree number A01, showing that it is directly under the category Anatomy [A] while MeSH descriptor "Mouth" with Unique ID D009055 has tree numbers: A01.456.505.631 (Body Regions → Head → Face → Mouth), A03.556.500 (Digestive System → Gastrointestinal Tract → Mouth) and A14.549 (Stomatognathic System → Mouth). A graphical representation of the MeSH hierarchy

for descriptor "Mouth" can be found at Figure 1.2.[1]

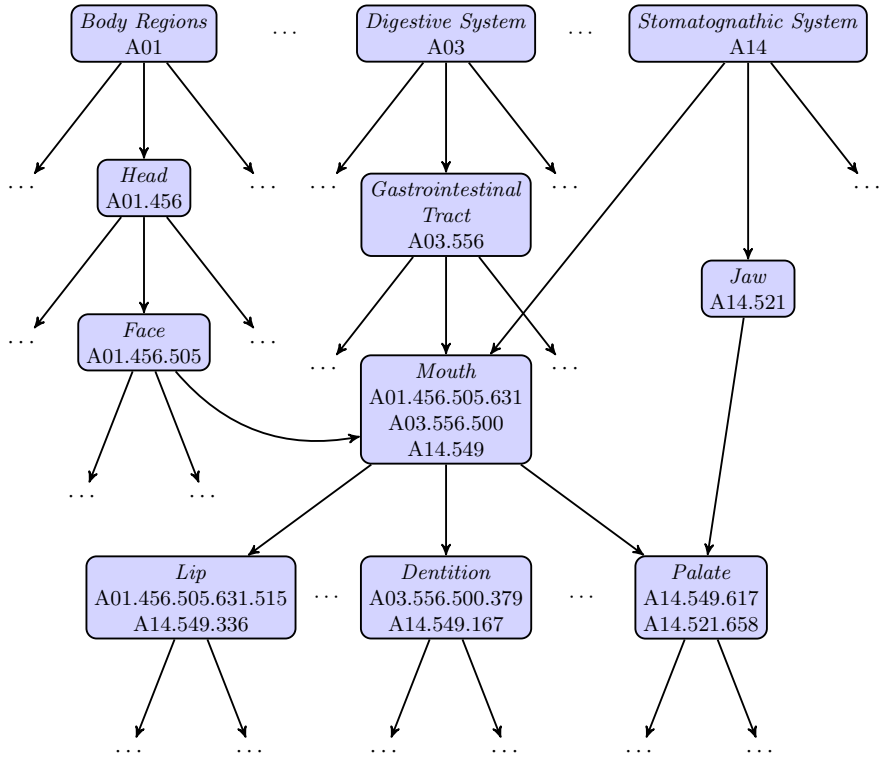


Figure 1.2: Partial graph of the MeSH hierarchy for descriptor "Mouth".

1.2.3 Qualifiers

Apart from the descriptors, which form the bulk of the MeSH catalogation system, there are 83 *topical qualifiers*, also known as *subheadings*, used for indexing and cataloging in conjunction with descriptors. This subheadings are not required but are used to put emphasis in an specialization of the given descriptor. Furthermore, each descriptor can only be qualified by a small set of subheadings. Suitable subheadings for the descriptor "Endocrine Cells" are, for example, "EN Enzymology" or "ME Metabolism".
It is worth noting that topical qualifiers have their own Unique Identifier (starting with Q and follow by 6 to 9 digits) and entry terms but do not belong to the hierarchy, therefore they have no tree number attached. As an example, qualifyer "ME Metabolism" has Unique ID Q000378 and entry terms "catabolysm",

[1]The entire tree structure can be browsed online on `nlm.nih.gov/mesh/2014/mesh_browser/MeSHtree.html`.

"biodegradation", among others.

1.2.4 PubMed articles

For the interest of this study we note that every article in the PubMed database has been manually annotated by domain experts with a set of MeSH terms [14].[2] Every publication is typically assigned between ten to twelve descriptors [15] and these data can be easily accessed via the PubMed web interface. Moreover, a subset of these descriptors is signaled as representing the major focus of the article, these are called the *major descriptors* of the article.[3] A PubMed article has on average four major descriptors.[4]
As an example, article "The effect of mepyramine and 48/80 on the histamine content of pleural exudates in the rat" (PubMed ID 999393) is paired with the descriptors "Animals", "Carrageenan", "Exudates and Transudates"(m), "Histamine"(m), "Pleura"(m), "Pleurisy", among others.
It is also worth noticing that every descriptor on an article can be assigned a subheading to refine the focus of the article but we will not consider them for this project.

[2]This denominations is used to loosely refer to MeSH descriptors.
[3]Some literature uses the term "major topics". We prefer to use the term "major descriptors" to avoid any confusion with LDA topics.
[4]Calculated on a sample of 100 000 articles.

Chapter 2

MeSH Correlation

In the first part of the project we worked with documents composed by the title and abstract of articles obtained from the PubMed database. The corpus contains 100 000 documents closely related to Neuroscience; every article has at least one MeSH descriptor under the category "Nervous System". The corpus was preprocessed using Bluima [18].

2.1 Preprocessing

The corpus went through minimal preprocesssing: sentence splitting and tokenization is done using OpenNLP, a machine learning toolkit for natural language processing; tokens are later lemmatized using BioLemmatizer [12], an open source lemmatization tool tailored to biomedical literature; finally, measure extraction, punctuation filtering and stopword filtering is performed. A stopword list with 524 general english words was used.
No frecuency filtering or further postprocessing was performed.

2.2 Topic Model

We trained the model using DCA with 200 topics and 500 full Gibbs cycles over the whole corpus; this took around 50 minutes running on 4 threads in a modern laptop (Intel® Core™ i5-3210M CPU @ 2.50GHz). From a qualitative point of view, the topic model did a good job at representing the documents in the corpus and showed its best performance in long technical abstracts with specific words and not crowded with quantities.
For a given document we define the concept of *Major Topics* as those that have a probability higher than a given threshold (0.05 in this study) and use the concept of *Major Descriptors* as described in Section 1.2. In this specific corpus, every document has an average of 4 major descriptors and 5 major topics.

To prove the suitability of the topic model we analyzed various articles chosen at random. Here we show the conclusions reached for two of them[1].

2.2.1 Analysis of Article 10996818

Heart rate measures in blind cave crayfish during environmental disturbances and social interactions.

Most animals continually assess the environment in which they live and alter their behavior according to various stimuli. As an observer, one looks for changes in a behavior indicating that an animal responded to a particular event. When the animal does not make significant behavioral changes as measured by bodily movements, the animal may be characterized as unresponsive to a given stimulus. This study demonstrates that when behavioral body movements can not be observed an internal physiological measure of heart rate (HR) shows dramatic changes following presentation of defined stimuli. This study used the blind cave crayfish and examined their responsiveness to the following stimuli: light (infrared, dim red, and white), water-borne vibrations, removal of water, olfactory cues, and social interaction with partners. This study demonstrates that there is substantial individual variation of HR at basal levels and with the intensity of an social interaction. We find HR is a reasonable measure of the responsiveness of blind cave crayfish to given stimuli even in the absence of observable behavioral changes. This enables the observer to determine if an individual is responsive to and making an assessment of particular cues.

MeSH descriptors(10):

Animals	Light	**Social Behavior**
Astacoidea	Male	Vibration
Environment	Olfactory Pathways	
Heart Rate	Physical Stimulation	

Figure 2.1: Top: Abstract for article PMID 10996818. Bottom: List of MeSH descriptors (major descriptors are shown in bold).

Article 10996818 is related to behavioral studies on a species of crayfish and has been written using specific but common language, avoiding technical terms. It is a relatively long abstract compared to the corpus used. Descriptors are well assigned and major descriptors are carefully chosen to represent the major focus of the article[2].

[1] The complete information of a PubMed article can be accesed via `http://www.ncbi.nlm.nih.gov/pubmed/?term=PMID`, where PMID is the PubMed ID.

[2] Astacoidea is a synonym for crayfish.

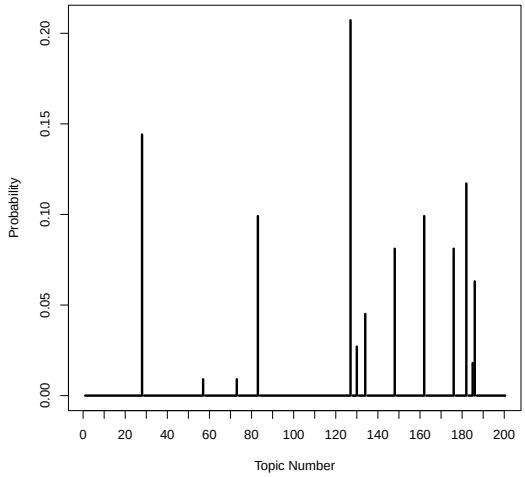


Figure 2.2: Topic distribution for article PMID 10996818 (see Figure 2.1). Most probable topics are 126, 27 and 181 (see Table 2.1).

Its topic distribution is given in Figure 2.2. This document is related to 13 different topics out of which 8 are major topics and 2 have a negligible probability ($p(z|d) < 0.01$) indicating a relatively flat topic distribution. This is a product of the general language used in the abstract.

In Table 2.1 we list all topics related to this document and the 10 most probable words for each one of them. The first major topic (126) represents very general language and is, in fact, the most common topic in the entire corpus. We observe that topic 82 and 161 have a high semantic relation with the abstract even though they have low probabilities ($p(z = 82|d = 10996818) = 0.1$ and $p(z = 161|d = 10996818) = 0.1$) which points towards the fact that middle level topics could carry more descriptive content than the highest probability topics (often related to general words).

This last remark is a result of the "general language" topics generated by the standard topic model. Topics representing general English language are a common problem in the standard topic model and our model also suffers from it; these topics are usually disregarded or discarded. We keep them for the rest of our analysis in this section.

Finally, we notice that there is a good correlation between the descriptors and topics of the article: topic 181, 82 and 161 could be sensibly linked to descriptors "Social Behaviour", "Heart Rate" and "Astacoidea" respectively, which are, in fact, the three major descriptors of this article.

Topic 126	**0.207**	**Topic 27**	**0.144**	**Topic 181**	**0.117**	**Topic 82**	**0.099**	**Topic 161**	**0.099**
study	0.038	amygdala	0.035	emotional	0.040	pressure	0.051	fish	0.041
result	0.023	conditioned	0.027	social	0.035	blood	0.032	specie	0.024
effect	0.021	conditioning	0.027	amygdala	0.026	sympathetic	0.030	vertebrate	0.016
show	0.019	fear	0.025	emotion	0.021	arterial	0.021	brain	0.012
suggest	0.018	stimulus	0.021	migraine	0.018	MEASURE_	0.021	teleost	0.011
human	0.015	lesion	0.017	behavior	0.016	increase	0.021	evolution	0.010
change	0.014	response	0.013	facial	0.011	response	0.019	mammal	0.010
present	0.014	rat	0.011	response	0.011	heart	0.015	trout	0.010
increase	0.011	nucleus	0.011	aggression	0.011	rate	0.014	zebrafish	0.009
previous	0.011	cs	0.011	affective	0.010	cardiovascular	0.014	sea	0.008
Topic 147	**0.081**	**Topic 175**	**0.081**	**Topic 185**	**0.063**	**Topic 133**	**0.045**	**Topic 129**	**0.027**
response	0.125	rat	0.085	visual	0.070	peptide	0.016	patient	0.037
stimulus	0.042	increase	0.040	motion	0.022	insect	0.015	stimulation	0.035
stimulation	0.026	effect	0.029	field	0.021	ganglion	0.012	pd	0.035
frequency	0.016	significant	0.025	cortex	0.016	hormone	0.011	disease	0.030
MEASURE_	0.016	decrease	0.023	stimulus	0.016	cockroach	0.009	's	0.030
change	0.013	level	0.023	area	0.014	jh	0.007	parkinson	0.028
increase	0.013	change	0.023	orientation	0.013	crustacean	0.007	tremor	0.021
evoke	0.011	animal	0.022	spatial	0.012	abdominal	0.007	stn	0.019
amplitude	0.010	treatment	0.021	receptive	0.010	pyloric	0.007	subthalamic	0.016
signal	0.009	control	0.018	v1	0.010	corpus	0.007	dystonia	0.015

Topic 184	**0.018**	**Topic 56**	**0.009**	**Topic 72**	**0.009**
light	0.031	insulin	0.051	frog	0.036
mc	0.016	diet	0.034	xenopus	0.027
phytochrome	0.014	weight	0.025	bee	0.017
plant	0.013	body	0.024	amphibian	0.017
blue	0.012	glucose	0.024	laevis	0.015
photoreceptor	0.011	fat	0.019	rana	0.010
red	0.010	plasma	0.014	tadpole	0.010
cryptochrome	0.009	hypoglycemia	0.012	melanotrope	0.010
arabidopsis	0.008	obesity	0.012	bc	0.008
phyb	0.008	dietary	0.011	honeybee	0.008

Table 2.1: Topics related to article PMID 10996818 ordered by $p(z|d)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

2.2.2 Analysis of Article 13130504

The abstract and MeSH descriptors for article 13130504 can be found at Figure 2.3. The article talks about the role of protein kinase B (Akt1) in the maintenance of cellular integrity during neuronal injuries. It uses specific technical language and is difficult to understand in its entirety without specific knowledge of the subject. Descriptors and major descriptors appear to be well assigned.

This document is related to 8 different topics including 4 major topics and 3 topics with a negligible probability ($p(z|d) < 0.01$). This is a desirable topic distribution which relates the article to only a few number of topics; this fact is represented by the small number of peaks in Figure 2.4.
In Table 2.2 we list all the topics related to the document and the 10 most probable words for each one. The first major topic (60) concerns cell death and seems to be a sensible major topic for this abstract; however, it has a remarkably high probability ($p(z = 60|d = 13130504) = 0.69$) implying that this abstract could be represented on a 69% solely by this topic. A possible explanation for this unusually high probability is that the scientific terms used in the abstract

Akt1 protects against inflammatory microglial activation through maintenance of membrane asymmetry and modulation of cysteine protease activity.

In several cell systems, protein kinase B (Akt1) can promote cell growth and development, but the "antiapoptotic" pathways of this kinase that may offer protection against cellular inflammatory demise have not been defined. Given that early cellular membrane phosphatidylserine exposure is a critical component of apoptosis, we investigated the role of Akt1 during neuronal apoptotic injury. By employing differentiated SH-SY5Y neuronal cells that overexpress a constitutively active form of Akt1 (myristoylated Akt1), free radical-induced cell injury was assessed through trypan blue dye exclusion, DNA fragmentation, membrane phosphatidylserine exposure, protein kinase B phosphorylation, cysteine protease activity, and mitochondrial membrane potential. Membrane phosphatidylserine exposure was both necessary and sufficient for microglial activation, insofar as cotreatment with an antiphosphatidylserine receptor-neutralizing antibody could prevent microglial activity following neuronal loss of membrane asymmetry. Furthermore, expression of myristoylated Akt1 not only prevented cell injury through the prevention of membrane phosphatidylserine exposure and genomic DNA fragmentation but also inhibited microglial activation and proliferation that required the inhibition of caspase 9-, caspase 3-, and caspase 1-like activities linked to cytochrome c release. Interestingly, Akt1 modulation of membrane phosphatidylserine exposure was primarily through caspase 1 activity. Removal of Akt1 activity abolished neuronal protection, suggesting that Akt1 functions as a critical pathway for the maintenance of cellular integrity and the prevention of phagocytic cellular removal during neurodegenerative insults.

MeSH descriptors(12):

Cell Membrane	Inflammation
Cell Survival	**Microglia**
Cysteine Endopeptidases	**Protein-Serine-Threonine Kinases**
Cytoprotection	**Proto-Oncogene Proteins**
Enzyme Activation	Proto-Oncogene Proteins c-akt
Humans	Tumor Cells, Cultured

Figure 2.3: Top: Abstract for article PMID 13130504. Bottom: List of MeSH descriptors (major descriptors are shown in bold).

may only be present on topic 60. Nonetheless, the other major topics also carry high expressive content: topic 163 ("protein kinases") and 65 ("microgial cells") are very specific to this article and various of their words even appear on the title of the article.

As in the previous example, we observe good results regarding the correlation

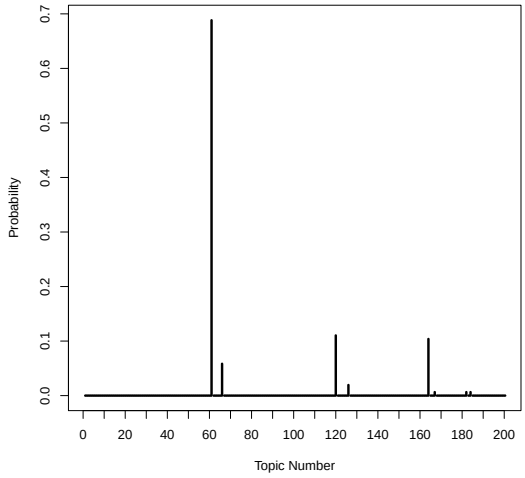


Figure 2.4: Topic distribution for article PMID 13130504 (see Figure 2.3). Most probable topics are 60, 119 and 163 (see Table 2.2).

Topic 60	**0.688**	**Topic 119**	**0.110**	**Topic 163**	**0.104**	**Topic 65**	**0.058**	**Topic 125**	**0.019**
cell	0.058	role	0.032	kinase	0.062	microglia	0.088	factor	0.087
death	0.056	function	0.022	protein	0.060	microglial	0.066	growth	0.054
apoptosis	0.040	mechanism	0.021	phosphorylation	0.037	cell	0.037	bdnf	0.048
neuronal	0.032	system	0.020	pkc	0.024	macrophage	0.033	neurotrophic	0.040
neuron	0.024	pathway	0.015	activity	0.023	activation	0.030	ngf	0.029
induce	0.017	play	0.014	camp	0.020	activate	0.021	neurotrophin	0.023
apoptotic	0.016	neuronal	0.013	inhibitor	0.016	complement	0.014	receptor	0.022
activation	0.012	involve	0.012	increase	0.014	inflammatory	0.011	gdnf	0.021
dna	0.011	important	0.012	activation	0.014	phagocytosis	0.009	survival	0.017
culture	0.011	regulate	0.011	pka	0.013	brain	0.008	derived	0.017

Topic 166	**0.006**	**Topic 181**	**0.006**	**Topic 183**	**0.006**
system	0.232	emotional	0.040	MEASURE_	0.661
central	0.175	social	0.035	al.	0.014
nervous	0.167	amygdala	0.026	j.	0.010
cns	0.081	emotion	0.021	respective	0.009
peripheral	0.054	migraine	0.018	study	0.009
tissue	0.008	behavior	0.016	micro	0.008
pns	0.004	facial	0.011	previous	0.005
organ	0.004	response	0.011	hr	0.005
function	0.004	aggression	0.011	type	0.005
periphery	0.004	affective	0.010	find	0.004

Table 2.2: Topics related to article PMID 13130504 ordered by $p(z|d)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

between MeSH descriptors and topics, e.g., topic 163 and descriptor "Protein-Serine-Threonine Kinases" or topic 65 and descriptor "Microglia". Despite the fact that topic 60 ("cell death") does not appear to have a strong relation with

the descriptors in the list we have to notice that descriptors "Cell Survival" and "Cell Death" are actually under the same branch of the MeSH hierarchy ("Cell Physiological Processes") and therefore topic 60 could be linked to descriptor "Cell Survival" and to the article.

2.3 MeSH Correlation

The probability of co-ocurrence of a MeSH descriptor and a topic on the corpus is given by:

$$p(m,z) = \sum_{d \in \mathcal{C}} p(m,z|d)p(d)$$

where d is a document, $\mathcal{C}$ is the corpus, z is a topic and m is a MeSH descriptor. We define $p(m|d) = 1$ when descriptor m is present in document d and 0 otherwise. Under the assumption that $z \perp\!\!\!\perp m \,|\, d$ and that $p(d) = \frac{1}{|\mathcal{C}|}$ we obtain:

$$p(m,z) = \frac{1}{|\mathcal{C}|} \sum_{d \in D(m)} p(z|d) \tag{2.1}$$

where $D(m)$ is the set of documents where m is present.
When going through the summation on Equation 2.1 we can choose to consider all the descriptors of a document or only the major ones, thus effectively assigning $D(m)$ to be the set of documents where m is a major descriptor. Likewise, we can choose to consider all the topics of a document or only the major ones, effectively assigning $p(z|d) = 0$ to non major topics and renormalizing $p(z|d)$. Major descriptors and major topics for a document (abbreviated MD and MT) were described in Section 2.2.
We use this probability to generate a so called *correlation matrix* between topics and MeSH descriptors.[3]

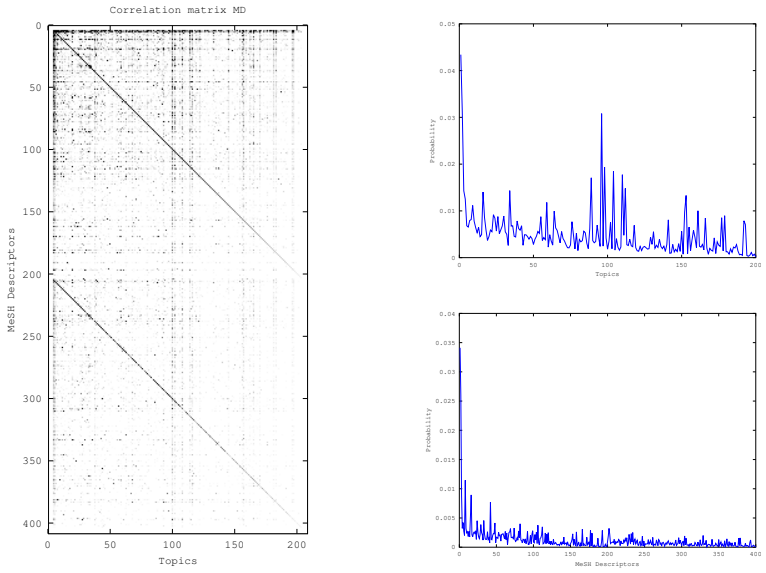
2.3.1 Visualization

Rows (MeSH descriptors) and columns (topics) on the correlation matrix could be reordered to reveal hidden structure in the data[4]. We enlist the different methods considered for this project accompanied by an example and a plot of $p(z) = \sum_m p(m,z)$ and $p(m) = \sum_z p(m,z)$ for the resulting matrix:

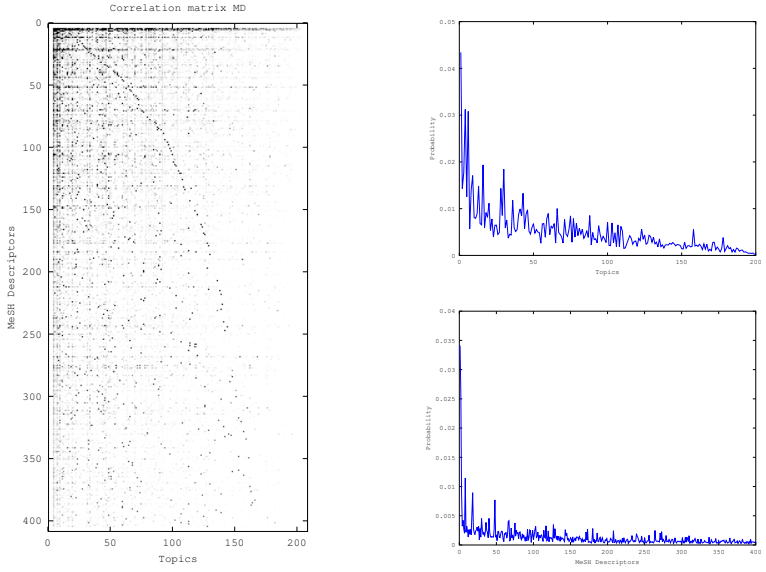
- Maximize the values on the diagonal of the matrix to set the first columns and rows. Once these are fixed, continue maximizing the diagonal to set the remaining rows.

[3]The term correlation is used throughout this section to refer to $p(m,z)$ unless stated otherwise.

[4]This is a research problem in itself generally referred as *seriation* or *sequencing*.

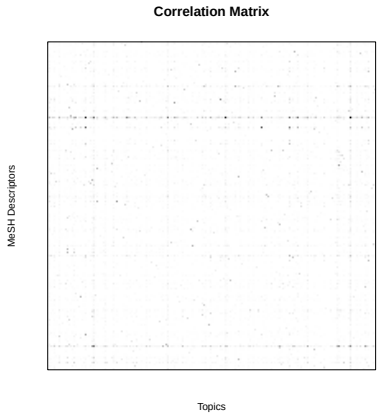


- Push high values to the upper left corner of the matrix such that the highest kth value is at worst in a k-sized box on the upper left corner.

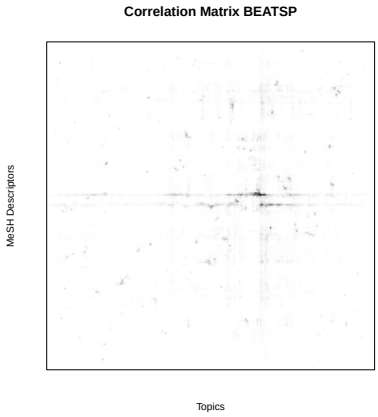


- Use the R seriation package [9] with its different heuristics to reorder the matrix. For these examples we used as the input a short version of our original matrix (200x200) because the package was not able to handle the

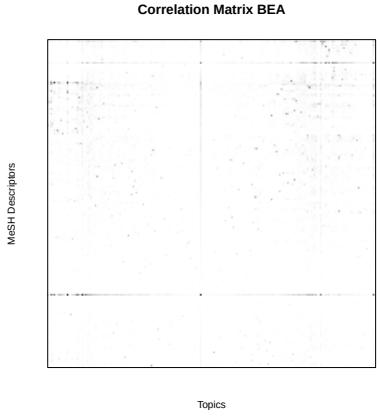
complete one (11642 x 200).



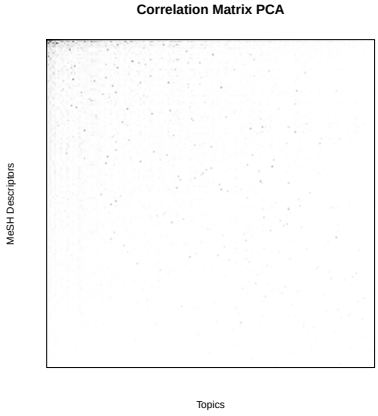
Input matrix



Seriation using BEA_TSP

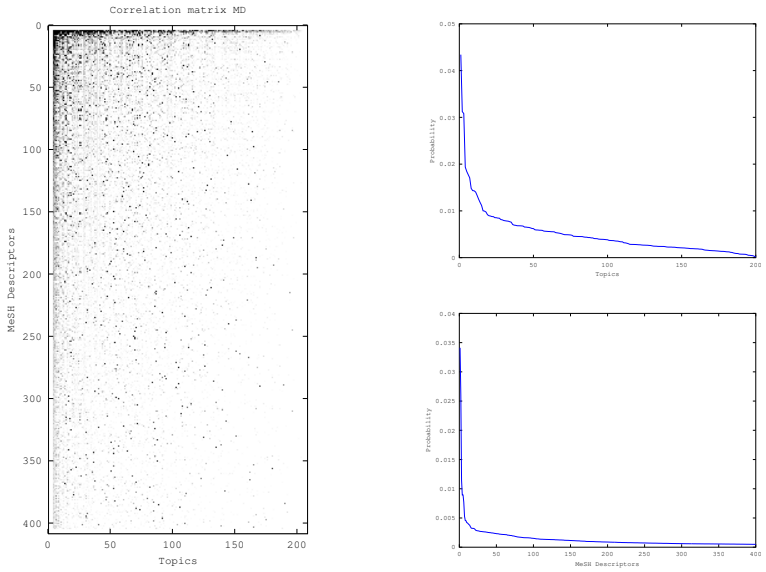


Seriation using BEA

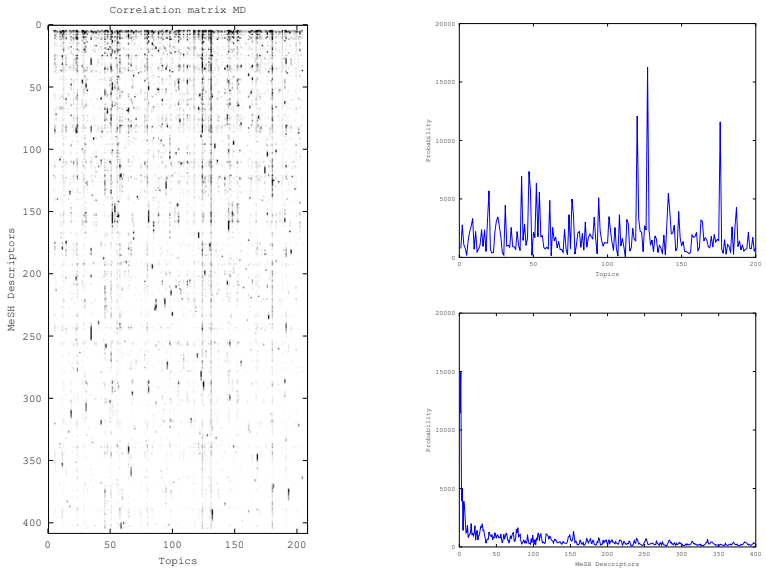


Seriation using PCA

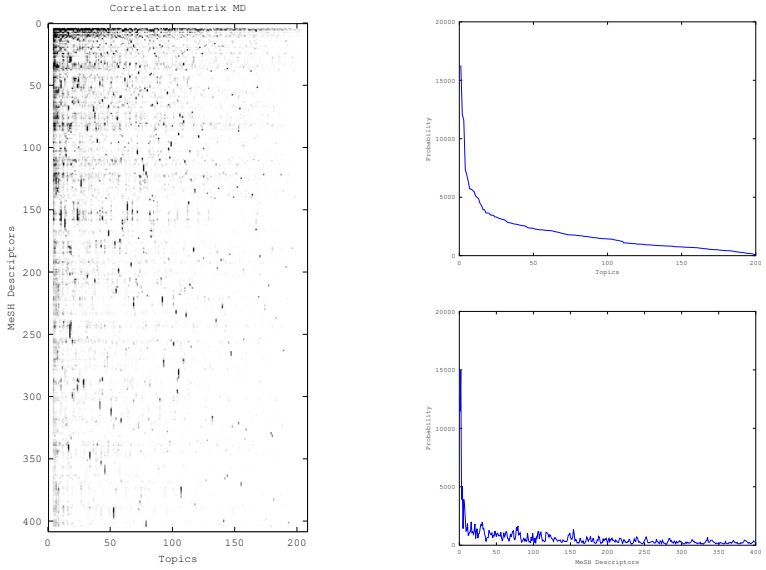
- Order columns on descending order defining the value of a column as the sum over its rows, i.e., order the topics by $p(z)$. Once columns are fixed order rows in a similar fashion.



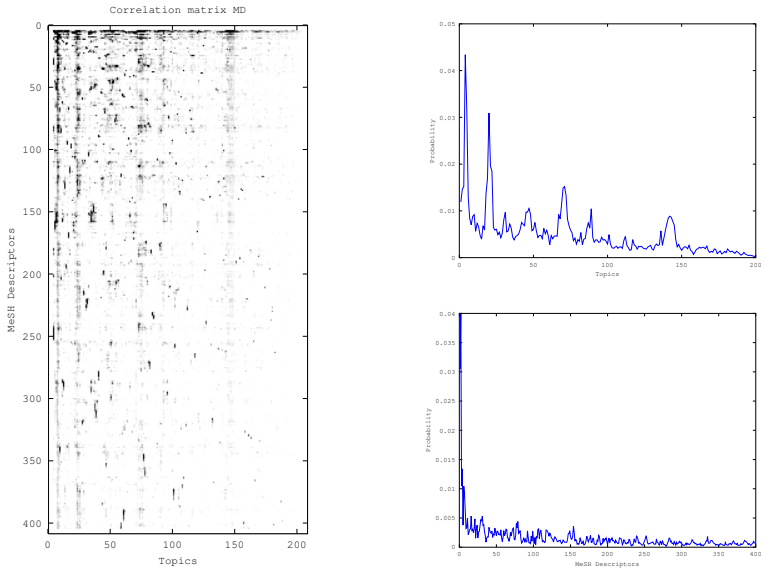
- Set the last row of the matrix to be the row with lowest probability ($p(m)$). Build the matrix from the bottom up by adding the row from the remaining set that is closer to the one placed last (considering each row as a vector and using the euclidean distance as a similarity measure). Columns are not reordered.



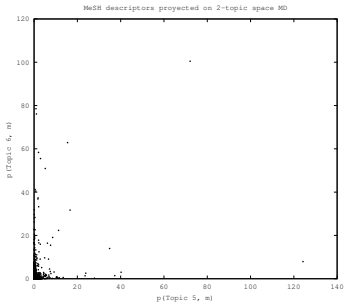
- Set the last row of the matrix to be the row with lowest probability ($p(m)$). Build the matrix from the bottom up by adding the row from the remaining set that is closer to the one placed last (considering each row as a vector and using the euclidean distance as a similarity measure). Once rows are set, order columns in descending order by $p(z)$.



- Set the last row of the matrix to be the row with lowest probability ($p(m)$). Build the matrix from the bottom up by adding the row from the remaining set that is closer to the one placed last (considering each row as a vector and using the euclidean distance as a similarity measure). Once rows are fixed, reorder columns in a similar fashion.



- Rather than presenting the data itself (correlation matrix) it could be useful to present some statistics derived from it. For this purpose, we plot each MeSH descriptor m as a point in a 2 dimensional plane whose axis are the $p(m, z)$ for a selected pair of topics z. As we are looking for descriptors that highly correlate with only a small subset of topics we expect each point in our graph to relate highly with only one of the topics (or to not relate with any of them). Thus, we expect curves similar to an exponential probability distribution.



We have chosen the first option to present our results given that it is simple, fast and good enough to show the relation between topics and descriptors.

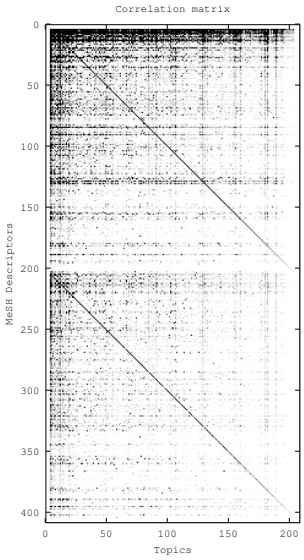
2.3.2 Results

We generated four different correlation matrices: one using all descriptors and all topics, one using all descriptors but only major topics, one using only major

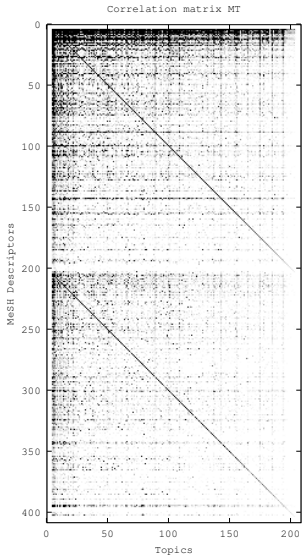
descriptors and all topics and one using only major descriptors and only major topics.
Results can be found on Figure 2.5. Every matrix has been reordered such as to maximize the value in the diagonal as explained in Section 2.3.1.
We can observe a clear line in the diagonal of our matrices indicating that there are various topic-descriptor pairs with a high correlation value. Certain general topics such as topic 126 shown in Table 2.1 have a high correlation with the majority of descriptors; this fact is illustrated in the matrices by dark vertical lines. Likewise, general descriptors such as "Brain" relate to a big proportion of topics, which is shown in our matrices by dark horizontal lines.
We are interested on matrices where descriptors relate strongly to only a few number of topics, i.e, a matrix with high values on the diagonal and some few other spots but low values elsewhere. This kind of matrix allows us to identify each MeSH descriptor by a probability distribution over a small set of topics.
Matrices MD and MD MT appear to show cleaner results, pointing towards the observation that selecting only major descriptors to compute this type of correlation yields better results than selecting all of them. Using only major descriptors is a sensible choice given that these have been handpicked by experts to express the major focus of the articles. On the other hand, major topics are assigned by the model and may not be descriptive enough given the generalization problems noted before, i.e., some major topics might be very general topics and not specific enough for each article while non-major topics might still carry important information. Altogether we believe the best outcome is obtained on matrix MD and proceed to analyze some of its features on the next section.

Analysis of Correlation Matrix MD

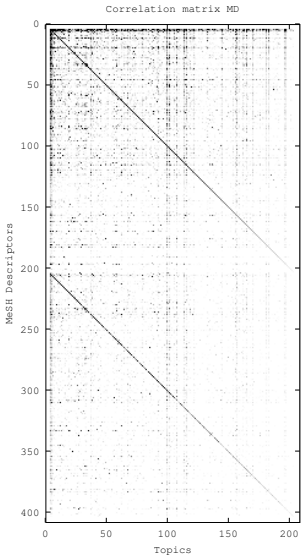
Less than half of all MeSH descriptors appear in our corpus (11641 out of 27149) and only 9000 descriptors appear in at least 2 documents. Figure 2.6 shows the number of descriptors appearing in different number of documents (plotted from 1-100 document appearances). These results come from the fact that our corpus is relatively small and focused in Neuroscience thus some MeSH descriptors are less probable to appear in the selected documents. We expect these numbers to improve as we move to a bigger corpus.
To verify that our results are semantically valid we performed a qualitative analysis similar to that performed when verifying a topic model: we show for each MeSH descriptor a list of topics to which it relates (ordered by $p(z|m)$) and the information for each topic.
In Table 2.3 we present results for the most frequent MeSH descriptor in our corpus (appearing in 14560 documents): descriptor "Brain". We observe that the highest related topic (126) is also the most general topic in our corpus; an expected result given that topic 126 will probably relate highly to the majority of the descriptors (although we do not expect it to be the highest in all of them). We can also observe that even though some of the topics have the word "brain" in them they could still be considered general topics, which is again expected given that the descriptor itself is not very specific. Finally, we note



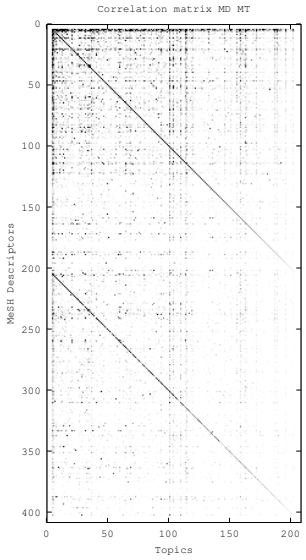
(a) All descriptors and all topics



(b) All descriptors and major topics



(c) Major descriptors and all topics



(d) Major topics and major descriptors

Figure 2.5: Correlation matrices for the four different descriptor and topics combinations used to calculate $p(m, z)$ on the 100K corpus (as explained on Section 2.3.2). Reordered as to maximize the diagonal values and cropped to match twice the number of topics (400)

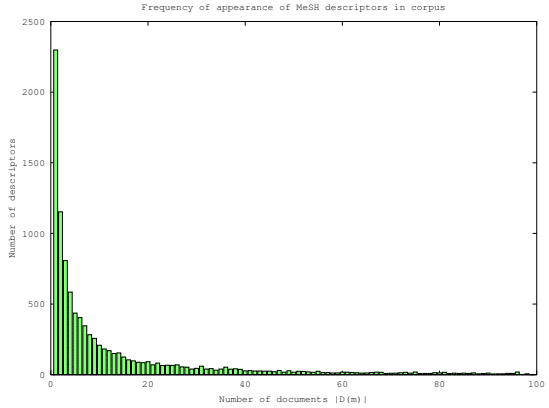


Figure 2.6: Frequency of appearances of MeSH descriptors in 100K corpus. Each bar represents the number of descriptors that appear in x documents (in x axis).

Topic 126	**0.045**	**Topic 140**	**0.039**	**Topic 51**	**0.035**	**Topic 175**	**0.030**	**Topic 119**	**0.030**
study	0.038	method	0.031	study	0.019	rat	0.085	role	0.032
result	0.023	datum	0.018	brain	0.016	increase	0.040	function	0.022
effect	0.021	image	0.016	review	0.015	effect	0.029	mechanism	0.021
show	0.019	analysis	0.016	research	0.011	significant	0.025	system	0.020
suggest	0.018	brain	0.011	human	0.011	decrease	0.023	pathway	0.015
human	0.015	technique	0.010	model	0.011	level	0.023	play	0.014
change	0.014	estimate	0.007	recent	0.010	change	0.023	neuronal	0.013
present	0.014	model	0.007	provide	0.010	animal	0.022	involve	0.012
increase	0.011	approach	0.007	development	0.009	treatment	0.021	important	0.012
previous	0.011	result	0.007	clinical	0.009	control	0.018	regulate	0.011

Table 2.3: Five highest related topics for descriptor "Brain" ordered by $p(z|m)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

that mostly all the topics presented have the same low $p(z|m)$, i.e., $p(z|m)$ is a very flat distribution, which is not the kind of distribution we are looking for. In summary, being "Brain" the most frequent MeSH descriptor it ends up picking the most general features of our corpus and relating to the most general topics of our model.

In Table 2.4 we present results for the 10^{th} most frequent descriptor in our corpus (appearing in 1849 documents): "Alzheimer Disease". This is a more specific (and interesting) descriptor and we see that the two highest correlated topics show a clear conceptual link to alzheimer disease. The other three topics could be considered general topics (given our corpus) but are not completely unrelated to alzheimer disease and have lower probabilities $p(z|m)$ compared with the first two.

In general, common descriptors and topics have higher correlation values when compared with less common descriptors and topics, which does not imply that we will consistently produce bad results when dealing with common descriptors. In fact, as long as the topic model generates only a small number of

Topic 144	**0.163**	**Topic 79**	**0.134**	**Topic 126**	**0.051**	**Topic 119**	**0.034**	**Topic 51**	**0.033**
disease	0.070	abeta	0.047	study	0.038	role	0.032	study	0.019
's	0.052	amyloid	0.042	result	0.023	function	0.022	brain	0.016
ad	0.045	ad	0.041	effect	0.021	mechanism	0.021	review	0.015
alzheimer	0.040	alzheimer	0.037	show	0.019	system	0.020	research	0.011
tau	0.032	disease	0.036	suggest	0.018	pathway	0.015	human	0.011
dementia	0.025	's	0.034	human	0.015	play	0.014	model	0.011
patient	0.018	plaque	0.021	change	0.014	neuronal	0.013	recent	0.010
alpha-synuclein	0.013	protein	0.020	present	0.014	involve	0.012	provide	0.010
body	0.012	peptide	0.020	increase	0.011	important	0.012	development	0.009
brain	0.011	app	0.019	previous	0.011	regulate	0.011	clinical	0.009

Table 2.4: Five highest related topics for descriptor "Alzheimer Disease" ordered by $p(z|m)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

general topics and all descriptors appear relatively often in the corpus results will be positive. Nonetheless, we will not be able to completely remove this effect given that our corpus will certainly have common descriptors and our model will certainly produce general topics.

Topic 26	**0.156**	**Topic 126**	**0.079**	**Topic 100**	**0.057**	**Topic 23**	**0.057**	**Topic 185**	**0.049**
movement	0.063	study	0.038	neural	0.026	model	0.068	visual	0.070
motor	0.057	result	0.023	network	0.024	parameter	0.012	motion	0.022
hand	0.021	effect	0.021	information	0.018	time	0.009	field	0.021
control	0.017	show	0.019	system	0.015	dynamics	0.009	cortex	0.016
task	0.015	suggest	0.018	sensory	0.014	simulation	0.009	stimulus	0.016
finger	0.013	human	0.015	processing	0.011	experimental	0.008	area	0.014
subject	0.012	change	0.014	circuit	0.010	rate	0.008	orientation	0.013
force	0.011	present	0.014	model	0.009	noise	0.008	spatial	0.012
limb	0.009	increase	0.011	brain	0.009	predict	0.008	receptive	0.010
arm	0.008	previous	0.011	pattern	0.008	datum	0.007	v1	0.010

Table 2.5: Five highest related topics for descriptor "Feedback" ordered by $p(z|m)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

We continue showing the results for the 1000^{th} most frequent descriptor in our corpus (appearing in 79 documents): "Feedback". This is a MeSH descriptor under the category of "Information Science". Even though the descriptor has a good number of appearances in the corpus, results do not appear to be very good at first inspection; this could be because the corpus was selected to relate to "Nervous System" and therefore the topics generated intrinsically relate to Neuroscience more than to Information Science. Taking this into account, topic 26, 100 and 23 which all relate to experimentation could be assumed to be a good correlation to "Feedback" given our corpus.
We present the results for the 5000^{th} most frequent descriptor in our corpus (appearing in 7 documents): "Neuropilin-1".[5] As expected descriptors become more specific as their frequency declines but even though "Neuropilin-1" only appears in seven documents it should still offer good results.

Topic 10 and 3 directly relate to the two primary functions of neuropilin-1:

[5] A protein involved in vessel formation and axonal guidance.

Topic 10	**0.357**	**Topic 30**	**0.077**	**Topic 3**	**0.074**	**Topic 38**	**0.059**	**Topic 119**	**0.049**
axon	0.046	expression	0.027	endothelial	0.067	nerve	0.092	role	0.032
growth	0.030	neural	0.023	vascular	0.046	regeneration	0.031	function	0.022
neurite	0.028	gene	0.021	vessel	0.031	axon	0.029	mechanism	0.021
outgrowth	0.019	development	0.019	cell	0.029	injury	0.023	system	0.020
adhesion	0.017	embryo	0.016	vegf	0.023	axonal	0.020	pathway	0.015
cone	0.016	cell	0.012	capillary	0.019	peripheral	0.017	play	0.014
molecule	0.016	express	0.011	brain	0.019	sciatic	0.017	neuronal	0.013
cell	0.015	early	0.009	cerebral	0.018	schwann	0.014	involve	0.012
guidance	0.014	zebrafish	0.009	blood	0.018	lesion	0.013	important	0.012
axonal	0.012	develop	0.008	factor	0.016	fiber	0.011	regulate	0.011

Table 2.6: Five highest related topics for descriptor "Neuropilin-1" ordered by $p(z|m)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

axonal guidance and vascular formation. The other three topics are more general topics but still relate indirectly to the descriptor. These are particularly good results given the small sample of documents and the specificity of the concept. Assuming that results for a given descriptor improve with its number of appearances in the corpus we could infer that even with this small corpus we have generated at least 5000 good correlations.

Topic 119	**0.075**	**Topic 139**	**0.074**	**Topic 159**	**0.067**	**Topic 87**	**0.053**	**Topic 150**	**0.051**
role	0.032	synaptic	0.031	synaptic	0.060	dendritic	0.043	synaptic	0.057
function	0.022	excitatory	0.027	ltp	0.049	dendrite	0.038	release	0.043
mechanism	0.021	postsynaptic	0.026	long-term	0.049	axon	0.036	vesicle	0.031
system	0.020	inhibitory	0.025	plasticity	0.043	terminal	0.032	presynaptic	0.028
pathway	0.015	neuron	0.020	potentiation	0.041	neuron	0.029	terminal	0.026
play	0.014	transmission	0.019	induction	0.021	synaptic	0.027	neuromuscular	0.025
neuronal	0.013	current	0.018	hippocampal	0.020	spine	0.022	synapsis	0.024
involve	0.012	inhibition	0.016	induce	0.018	synapsis	0.021	synapse	0.019
important	0.012	presynaptic	0.015	depression	0.017	cell	0.019	junction	0.019
regulate	0.011	amplitude	0.014	synapsis	0.016	contact	0.012	postsynaptic	0.017

Table 2.7: Five highest related topics for descriptor "Synapses" ordered by $p(z|m)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

Finally we have handpicked a descriptor that appears often in the corpus and has at least two tree numbers: 'Synapses", the 12^{th} most frequent descriptor in the corpus with 1756 document appearances[6]. The first topic is again a general topic given our corpus but all other topics are clearly related to "Synapses" with each topic focusing in a different synaptic feature: excitation and inhibition, plasticity, parts of the neuron and operation, respectively.
As a whole, these results are encouraging and show a clear semantic correlation between topics and descriptors, which is remarkable given that we did not perform any special tuning on the corpus or the model. We expect these results to get better as we refine the topic model and increase the size of the corpus.
It is also worth noticing that we obtained better results than previously presented [21].

[6]Handpicked to be used later in Section 2.5

2.3.3 Applications

The generated topic model and correlation matrix could be used for various practical tasks; we enlist some of them:

- Descriptor prediction for a document d could be performed by building a ranking system using $p(m|d)$.
- Prediction of major descriptors for a document d with a given set of descriptors $M(d)$ could be done in a similar fashion.
- New relations on the MeSH hierarchy could be retrieved using information measures such as the symmetric KL divergence between descriptors.

Some of this experiments were performed on a smaller corpus in [16].

2.4 MeSH Correlation with graph nodes

In this section we use our knowledge of the MeSH structure to compute $p(n, z)$, the probability of co-ocurrence of graph node n and topic z. Nodes in the structure are uniquely identified by a single descriptor and may have various tree numbers as explained on Section 1.2.2.
We start from the assumption that a given descriptor m relates to a document (and its topics) not only when the document is tagged with m but also when the document is tagged with any subconcept of m, i.e., a descriptor that appears under m in the MeSH hierarchy. We expect to improve the results obtained using simple correlation because we believe that documents which relate to a subconcept of m also relate to m and thus should be considered in the correlation.
To avoid the confusion with the correlation with descriptors where this assumption is not made we say that it is the graph node n identified by m that relates to a document and thus compute the correlation between topics and nodes.[7]
The probability of co-ocurrence of a graph node and a topic is given by:

$$p(n, z) = \sum_{d \in \mathcal{C}} p(n, z|d)p(d)$$

where d is a document, $\mathcal{C}$ is the corpus, z is a topic and n is a graph node.
We define $p(n|d) = 1$ if node n is present in document d and 0 otherwise. We say graph node n is present in document d if for some descriptor m in d, n is the node identified by m or $n \in \mathit{Above}(m)$ where $\mathit{Above}(m)$ is constructed using the tree numbers of the node identified by m. For instance, node "Palate" has two tree numbers: A14.549.617 and A14.521.658; therefore Above("Palate") =

[7] The name *graph node* comes from the fact that we make use of the MeSH structure as a graph to compute the probabilities.

{A14 "Stomatognathic System", A14.549 "Mouth", A14.521"Jaw"}.[8]
Under the assumption that $z \perp\!\!\!\perp n \,|\, d$ and that $p(d) = \frac{1}{|\mathcal{C}|}$ we obtain:

$$p(n, z) = \frac{1}{|\mathcal{C}|} \sum_{d \in D(n)} p(z|d) \tag{2.2}$$

where $D(n)$ is the set of documents where n is present.
We use this probability to construct correlation matrices between topics and nodes. As a node is uniquely identified by a MeSH descriptor our matrix will have rows representing MeSH descriptors and columns representing topics; similar to matrices in Section 2.3. Notice however that in this case MeSH descriptors that do not explicitly appear in the corpus could appear as a row in the matrix given that this descriptors may be in the set $Above(m)$ for a descriptor m that appears in the corpus. Therefore, we will potentially produce bigger matrices than when correlating topics with descriptors.
This fact could be useful when dealing with a corpus where not all MeSH descriptors are present or appear scarcely. Choosing this type of correlation will fill the gaps of descriptors that do not appear in the corpus but are higher in the MeSH hierarchy than those appearing. For example, if descriptor "Mouth" does not appear in the corpus but descriptor "Palate" or any other descriptor under "Mouth" does, then "Mouth" will appear in the results. However, it is more probable than a general descriptor, like "Mouth", appears in the corpus rather than an specific one, like "Palate" or "Parotid Gland" (both under "Mouth").

2.4.1 Results

As explained in Section 2.3.2 we could create four different correlation matrices based on the data used in the computation of $p(n, z)$. Based on the results in the previous section we only generated matrix MD, it is, we used all topics but only major descriptors. We decided to continue with this option because it showed the most promising results.
As can be seen in Figure 2.7 our matrix still show clear diagonal lines signaling a big number of highly correlated pairs but in this case we notice that the number of dark horizontal lines, i.e, the number of graph nodes that have high correlation with all topics, has significatively increased, which is a direct result of the formulation of Equation 2.2.
Less than half of all MeSH descriptors appear in our results (13543 out of 27149) with 11000 appearing in at least 2 documents. Figure 2.10 shows the number of nodes appearing in different number of documents (plotted from 1-100 document appearances). Unsurprisingly, the three most frequent nodes are "Nervous System A08", "Central Nervous System A08.186" and "Brain A08.186.211", given that "Brain" was the most frequent descriptor in the corpus and all documents related to node "Brain" will also relate to node "Nervous System" and "Central Nervous System" (which are above "Brain"). In fact, given that each document

[8] Nodes used in this example appear in Figure 1.2.

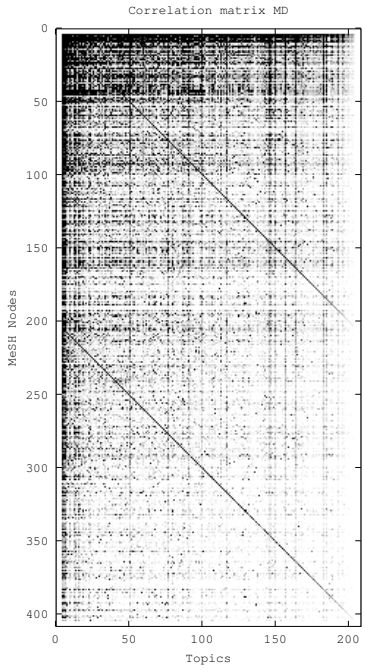


Figure 2.7: Correlation matrix using correlation with graph nodes in 100K. Reordered as to maximize the diagonal values and cropped to match twice the number of topics (400)

in the corpus was choosen to have at least one MeSH descriptor under "Nervous System", this node will appear in every document. The five highest correlated topics for these three nodes are a combination of a set of only seven topics out of which 126, 119 and 175 are shared by all; these are all very general topics and give us virtually no discerning power between these three descriptors. In this case, using correlation with graph nodes amplifies the problem of the general topics generated by our model and the common descriptors appearing in the data.

To compare it with the results obtained when using the simple correlation with descriptors we will inspect the results for the same descriptors used in the previous section.

In Table 2.8 we show the highest related topics for node "Brain". "Brain" is the third most frequent node in the corpus appearing in 48653 documents (compared to the 14954 documents it appeared as a descriptor). It is related with very general topics as before and values are relatively the same for all topics signaling a flat topic distribution. Results for this node do not show any improvement compared to the simple correlation with descriptors.

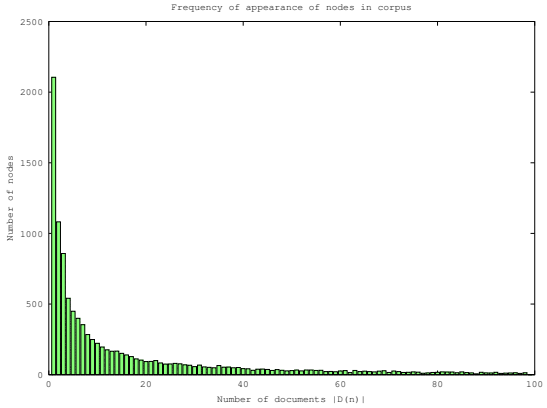


Figure 2.8: Frequency of appearances of graph nodes in 100K corpus. Each bar represents the number of nodes that appear in x documents (in x axis).

Topic 126	**0.045**	**Topic 175**	**0.034**	**Topic 119**	**0.026**	**Topic 47**	**0.025**	**Topic 140**	**0.019**
study	0.038	rat	0.085	role	0.032	activation	0.039	method	0.031
result	0.023	increase	0.040	function	0.022	cortex	0.037	datum	0.018
effect	0.021	effect	0.029	mechanism	0.021	area	0.029	image	0.016
show	0.019	significant	0.025	system	0.020	functional	0.027	analysis	0.016
suggest	0.018	decrease	0.023	pathway	0.015	region	0.025	brain	0.011
human	0.015	level	0.023	play	0.014	fmri	0.021	technique	0.010
change	0.014	change	0.023	neuronal	0.013	study	0.016	estimate	0.007
present	0.014	animal	0.022	involve	0.012	imaging	0.015	model	0.007
increase	0.011	treatment	0.021	important	0.012	brain	0.015	approach	0.007
previous	0.011	control	0.018	regulate	0.011	temporal	0.015	result	0.007

Table 2.8: Five highest related topics for node "Brain" ordered by $p(z|n)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

Descriptors "Alzheimer Disease" and "Neuropilin-1" appear as nodes in the same documents that they appear as descriptors and therefore have the same topic distribution.
Descriptor "Feedback" appears in 149 documents as a node (compared to the 79 documents it appeared as a descriptor) but the five highest topics it relates to are the same as before with the exception of topic 119 (related to vision) being replaced by topic 185 (related to function), which shows a slight improvement from the previous result although it could be a fortuitous event. Despite the fact that it relates to the same topics as before, the probabilities have slightly decreased meaning that this topic distribution is flatter than what we had when using only descriptors. Altogether, these remarks are not positive when related to the previous case.
We list the results for node "Synapses", which appears in 3266 documents as a node (compared to 1756 documents as a descriptor). Again the five highest topics it relates to are the same as in the previous version but we notice that the general topic 119 has been deplaced as the highest correlated topic by topic 150

Topic 26	**0.108**	**Topic 126**	**0.075**	**Topic 100**	**0.054**	**Topic 23**	**0.045**	**Topic 119**	**0.037**
movement	0.063	study	0.038	neural	0.026	model	0.068	role	0.032
motor	0.057	result	0.023	network	0.024	parameter	0.012	function	0.022
hand	0.021	effect	0.021	information	0.018	time	0.009	mechanism	0.021
control	0.017	show	0.019	system	0.015	dynamics	0.009	system	0.020
task	0.015	suggest	0.018	sensory	0.014	simulation	0.009	pathway	0.015
finger	0.013	human	0.015	processing	0.011	experimental	0.008	play	0.014
subject	0.012	change	0.014	circuit	0.010	rate	0.008	neuronal	0.013
force	0.011	present	0.014	model	0.009	noise	0.008	involve	0.012
limb	0.009	increase	0.011	brain	0.009	predict	0.008	important	0.012
arm	0.008	previous	0.011	pattern	0.008	datum	0.007	regulate	0.011

Table 2.9: Five highest related topics for node "Feedback" ordered by $p(z|n)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

Topic 150	**0.075**	**Topic 119**	**0.061**	**Topic 139**	**0.057**	**Topic 87**	**0.044**	**Topic 4**	**0.043**
synaptic	0.057	role	0.032	synaptic	0.031	dendritic	0.043	synaptic	0.060
release	0.043	function	0.022	excitatory	0.027	dendrite	0.038	ltp	0.049
vesicle	0.031	mechanism	0.021	postsynaptic	0.026	axon	0.036	long-term	0.049
presynaptic	0.028	system	0.020	inhibitory	0.025	terminal	0.032	plasticity	0.043
terminal	0.026	pathway	0.015	neuron	0.020	neuron	0.029	potentiation	0.041
neuromuscular	0.025	play	0.014	transmission	0.019	synaptic	0.027	induction	0.021
synapsis	0.024	neuronal	0.013	current	0.018	spine	0.022	hippocampal	0.020
synapse	0.019	involve	0.012	inhibition	0.016	synapsis	0.021	induce	0.018
junction	0.019	important	0.012	presynaptic	0.015	cell	0.019	depression	0.017
postsynaptic	0.017	regulate	0.011	amplitude	0.014	contact	0.012	synapsis	0.016

Table 2.10: Five highest related topics for node "Synapses" ordered by $p(z|n)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

(synaptic operation); this is a positive result albeit not very significative. The other topics have suffer some reordering but as they all relate to "Synapses" we believe the results are similar.

Even though the descriptors used in this analysis were chosen to be representative of those appearing in the corpus there is some concerns about the sample, for instance: descriptor "Brain" appears to be too general as to be a fair comparison; "Feedback", on the other hand, appears too specific; "Alzheimer Disease" and "Neuropilin-1" do not have any nodes under them in the MeSH hierarchy so there was no room for improvement when correlating them as nodes and "Synapses" could be think of as a best-case scenario for correlation. Nonetheless, if results were consistently better for any of the two methods we would have found some evidence in the data. Further examination should be conduced although a simple qualitative analysis may not be the ideal way to analyze the differences between these two correlations.

In summary, even after careful examination, results were not conclusive in favor of any method: we have shown that using correlation with graph nodes produces similar results to those produced by the correlation with descriptors. Although we found some positive results on using the graph nodes as part of the computation they were not significative or consistent in our sample. On the downside, using graph nodes seems to aggravate the problem of very general

topics and descriptors with flat topic distributions. This effect is corpus dependent: as we chose every document to be under descriptor "Nervous System" we expected "Nervous System" to be related with the most probable topics of the entire corpus, which is not a bad result in itself given that the corpus by definition relates to "Nervous System". Although the effects of using a bigger corpus remain to be seen, we believe this type of correlation should at least be an alternative to consider.

2.4.2 Aplications

Correlation using graph nodes can be used as an alternative to the simple correlation with descriptors in any of the applications listed in Section 2.3.3.
Additionally, correlation with graph nodes may be preferred over correlation with descriptors when we are more interested in capturing high level rather than specific correlations, for example, when building a recommender system that receives a document and generates a list of possible MeSH descriptors it could be more useful to propose higher level MeSH descriptors which may later be refined by a human tagger rather than specific MeSH descriptors which may be incorrect.

2.5 MeSH Correlation with tree numbers

2.5.1 Correlation with tree numbers

As explained in Section 1.2.2 a MeSH descriptor can appear in more than one place of the MeSH hierarchy, i.e., it can be classified in more than one way and each of this classifications is uniquely identified by a tree number. For instance, MeSH descriptor "Synapses" has two tree numbers: A08.850 (Nervous System → Synapses) and A11.284.149.165.420.780 (Cells → Cellular Structures → Cell Membrane → Cell Membrane Structures → Intercellular Junctions → Synapses).
We could use a derivation similar to the one used in Section 2.4 to correlate topics directly to tree numbers, it is, compute $p(t, z)$ where t is a tree number and z is a topic.
The probability of co-ocurrence of a tree number and a topic is given by:

$$p(t, z) = \sum_{d \in \mathcal{C}} p(t, z|d)p(d)$$

where d is a document, $\mathcal{C}$ is the corpus, z is a topic and t is a tree number.
We define $p(t|d) = 1$ if tree number t is present in document d and 0 otherwise. We say t is present in document d if for some descriptor m in d, t is a tree number of m or $t \in TAbove(m)$ where $TAbove(m)$ is constructed using the tree numbers of m. For instance, descriptor "Dentition" has two tree numbers: A03.556.500.379 and A14.549.167; therefore TAbove("Dentition") =

{A03, A03.556, A03.556.500, A14, A14.549}.[9]. Notice that different tree numbers for the same descriptor or even a subset of them could appear in the results as can be observed in the example with tree numbers A03.556.500 and A14.549 which both are tree numbers for descriptor "Mouth".
Under the assumption that $z \perp\!\!\!\perp t \,|\, d$ and that $p(d) = \frac{1}{|\mathcal{C}|}$ we obtain:

$$p(n, z) = \frac{1}{|\mathcal{C}|} \sum_{d \in D(t)} p(z|d) \tag{2.3}$$

where $D(t)$ is the set of documents where t is present.
Similarly to the case of correlation with graph nodes, tree numbers whose descriptor does not appear in the corpus may still appear in the results conceding that they are elements of the set $TAbove(m)$ for a descriptor m that appears in the corpus. Another effect of this definition is that for those descriptors with a single tree number t, the correlations assigned to t will indeed be exactly the same as those assigned to the descriptor in the correlation with graph nodes. Unlike the correlation with descriptors and the correlation with graph nodes, Equation 2.3 produces a correlation matrix with *tree numbers* as rows and topics as columns.
This kind of correlation could help to discriminate between different meanings of a descriptor; for example, one could argue that different tree numbers for the same MeSH descriptor refer to different concepts or have a particular focus and that this distinction can be captured by the relation between these different tree numbers and the topics.

2.5.2 Results

Figure 2.9 shows the results for matrix MD, i.e, $p(t, z)$ was computed using all topics but only major descriptors (as explained in Section 2.3.2). The correlation matrix looks very similar to that obtained in Section 2.4.1 with several dark horizontal and vertical lines which are not a positive indicator. Nevertheless, as in Section 2.4.1 it does not neccesarily mean that results are negative.
This agglomeration of high values in the first rows of the matrix may be produced by the fact that these tree numbers are *present* in a high number of documents in the corpus. In addition to that, as we saw in the last section, the topic distributions tend to become flatter when using this type of correlations which may also account for the dark horizontal lines. Dark vertical lines are a byproduct of the standard topic model as have been already documented.
There is 27591 tree numbers appearing in our corpus out of which around 27000 appear in at least two documents. The distribution of document appearances is presented on Figure 2.10. As in the correlation with graph nodes, the three more common tree numbers are the ones corresponding to descriptors "Nervous System", "Central Nervous System" and "Brain"; this descriptors have only one tree number and thus their topic distributions are exactly the same as before. Once again, this result was expected given the way our computation is

[9]Descriptors used in this example appear as a graph representation in Figure 1.2.

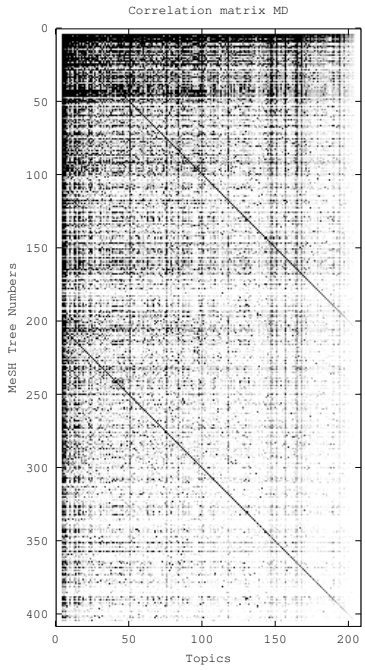


Figure 2.9: Correlation matrix using correlation with tree numbers in 100K. Reordered as to maximize the diagonal values and cropped to match twice the number of topics (400)

performed.
In order to evaluate the discerning power of this type of correlation we will analyze different tree numbers for a selected MeSH descriptor: "Synapses". We have chosen "Synapses" for various reasons: it appears often in the corpus, it has more than one tree number (shown below), it has been proven to generate good results when using our current topic model (as shown on Table 2.7) and is simple enough to actually differentiate between the two concepts without a high degree of expertise.
Descriptor "Synapses" has two tree numbers: A08.850 (Nervous System → Synapses) and A11.284.149.165.420.780 (Cells → Cellular Structures → Cell Membrane → Cell Membrane Structures → Intercellular Junctions → Synapses). Both tree numbers are under the category "Anatomy [A]" but A11.284.149.165.420.780 relates more specifically to Synapses as part of a cell while the former focuses in the anatomic concept of synapse.
A08.850 being more general appears in the documents a higher number of times (3266) than A11.284.149.165.420.780 (3005). We present the topic distribution for tree number A08.850 in Table 2.11 and for tree number A11.284.149.165.420.780

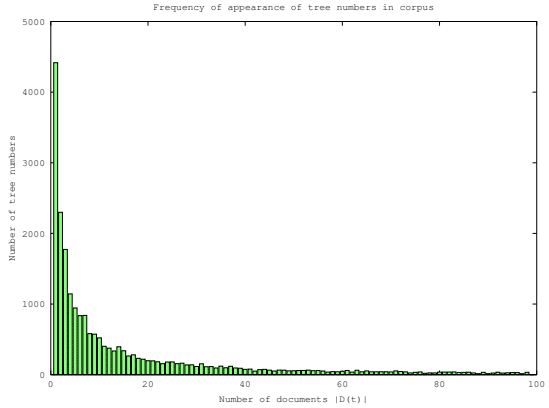


Figure 2.10: Frequency of appearances of tree numbers in 100K corpus. Each bar represents the number of tree numbers t that appear in x documents (in x axis).

in Table 2.12.

Topic 150	**0.075**	**Topic 119**	**0.061**	**Topic 139**	**0.057**	**Topic 87**	**0.044**	**Topic 4**	**0.043**
synaptic	0.057	role	0.032	synaptic	0.031	dendritic	0.043	synaptic	0.060
release	0.043	function	0.022	excitatory	0.027	dendrite	0.038	ltp	0.049
vesicle	0.031	mechanism	0.021	postsynaptic	0.026	axon	0.036	long-term	0.049
presynaptic	0.028	system	0.020	inhibitory	0.025	terminal	0.032	plasticity	0.043
terminal	0.026	pathway	0.015	neuron	0.020	neuron	0.029	potentiation	0.041
neuromuscular	0.025	play	0.014	transmission	0.019	synaptic	0.027	induction	0.021
synapsis	0.024	neuronal	0.013	current	0.018	spine	0.022	hippocampal	0.020
synapse	0.019	involve	0.012	inhibition	0.016	synapsis	0.021	induce	0.018
junction	0.019	important	0.012	presynaptic	0.015	cell	0.019	depression	0.017
postsynaptic	0.017	regulate	0.011	amplitude	0.014	contact	0.012	synapsis	0.016

Table 2.11: Five highest related topics for tree number A08.850 (Nervous System $\rightarrow$ Synapses) ordered by $p(z|t)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

The number of documents where both tree numbers are present highly overshadows the number of documents where only one of them is and therefore results are very similar for both; this is due to the fact that the set of tree numbers appearing under A08.850 (we'll refer to it as Nervous System for convenience) and the set of those appearing under A11.284.149.165.420.780 (we will call it Cell) are not distinct, in fact Cell is a proper subset of Nervous System with only one single MeSH descriptor("Synaptic Vesicles") which appears on Nervous System and does not appear on Cell. It is based on this single descriptor that we are trying to differentiate between the two tree numbers and therefore the two connotations of "Synapses".
It may seem that this was a poorly chosen example for this kind of correlation but in fact this relations are an intrinsical property of the MeSH hierarchy: gen-

Topic 150	**0.074**	**Topic 119**	**0.062**	**Topic 139**	**0.061**	**Topic 87**	**0.047**	**Topic 4**	**0.045**
synaptic	0.057	role	0.032	synaptic	0.031	dendritic	0.043	synaptic	0.060
release	0.043	function	0.022	excitatory	0.027	dendrite	0.038	ltp	0.049
vesicle	0.031	mechanism	0.021	postsynaptic	0.026	axon	0.036	long-term	0.049
presynaptic	0.028	system	0.020	inhibitory	0.025	terminal	0.032	plasticity	0.043
terminal	0.026	pathway	0.015	neuron	0.020	neuron	0.029	potentiation	0.041
neuromuscular	0.025	play	0.014	transmission	0.019	synaptic	0.027	induction	0.021
synapsis	0.024	neuronal	0.013	current	0.018	spine	0.022	hippocampal	0.020
synapse	0.019	involve	0.012	inhibition	0.016	synapsis	0.021	induce	0.018
junction	0.019	important	0.012	presynaptic	0.015	cell	0.019	depression	0.017
postsynaptic	0.017	regulate	0.011	amplitude	0.014	contact	0.012	synapsis	0.016

Table 2.12: Five highest related topics for tree number A11.284.149.165.420.780 (Cells → Cellular Structures → Cell Membrane → Cell Membrane Structures → Intercellular Junctions → Synapses) ordered by $p(z|t)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

erally when a descriptor appears in two different places in the hierarchy all the MeSH descriptors right below it also appear in both places and there is only a small number of descriptors that are not shared between these two branches. Even more, the subset relation is also preserved in these two branches. As a matter of fact descriptor "Synapses", being the 12^{th} most frequent descriptor in the corpus, could have been considered a best-case scenario for this type of correlation.
To further confirm our insights we performed the same experiment for descriptor "Neurons" with tree numbers A08.663 (Nervous System → Neurons) and A11.671 (Cell → Neurons). We obtained equally poor results: both topic distributions are esentially the same and the set of tree numbers for A11.671 is a proper subset of those for A08.663. In this case, A08.663 appeared 21698 times in the documents while A11.671 appeared in 21194 documents meaning that there is only 504 documents where A08.663 is present and A11.671 is not. Being A08.663 and A11.671 tree numbers found very high in the hierarchy (therefore potentially having a big amount of tree numbers under them) one could have expected to obtain a bigger difference between them.
This effects will not be attenuated when moving to a bigger corpus given that the proportion of documents where only one tree number appears will only scale but the end product will remain the same. In summary, unless a further examination of the MeSH structure is performed to reformulate Equation 2.3, this approach does not generate positive results mostly due to the intrinsic properties of the MeSH hierarchy.

2.5.3 Applications

Articles in the PubMed database are tagged only with MeSH descriptors but one could argue that adding the specific tree number for each MeSH descriptor is a valuable piece of information in refining the focus of the article. In this context, correlation with tree nodes could have various practical uses:

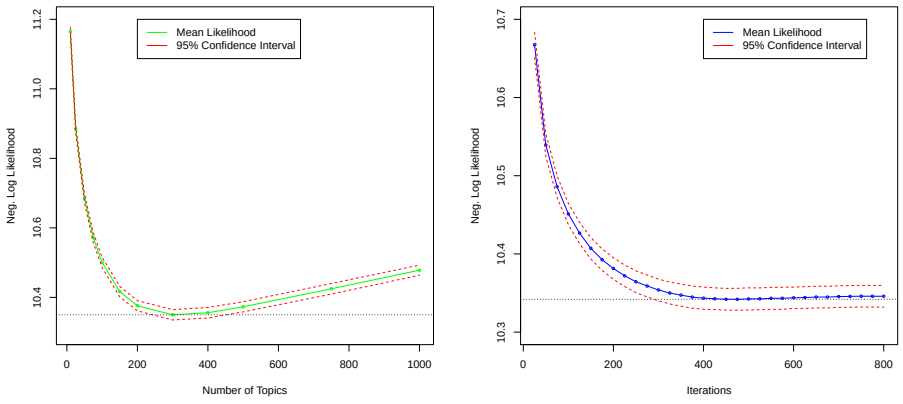
- As a prediction system that tags articles directly with tree numbers instead

of only MeSH descriptors.

- As an automatic tool that assigns the best tree number given an article already tagged with MeSH descriptors.
- As a metric to suggest the creation of a new MeSH descriptor when an existing descriptor has more than one concept attached to it.

2.6 Topic and iteration fitting

We ran topic and iteration fitting experiments on the corpus using 10 fold cross-validation. Likelihood on held-out data was calculated using Left-to-Right sequential sampler[5] as implemented in the DCA package; this algorithm is shown to be unbiased on large corpus.



(a) Topic fitting using 500 iterations

(b) Iteration fitting using 300 topics

Figure 2.11: Topic and iteration fitting results using 10-fold cross validation on 100K corpus.

As observed on 2.11a, the negative log likelihood decreases rapidly while the number of topics increases but slows around 300 topics and starts increasing again at 400, effect probably produced by overfitting.
The likelihood for iterations show an steady decrease before reaching 500 iterations where it stops decreasing and slightly raises again, although not significatively.

2.7 Proposed changes

There are various paths we could take to improve the quality of our topic model.

- Improve the fitting of topics on the DCA model: using crossvalidation on the corpus to obtain the most suitable number of topics and number of iterations. As performed in Section 2.6
- Improve the already generated model: identifying "general purpose" topics and incoherent topics
- Improve preprocessing:
  - Improve the lemmatization of tokens.
  - Use a preexisting vocabulary in the tokenization process.
  - Add new words to the stopword list.
  - Add multiword tokenization to the preprocesing chain.
- Improve visualization of results:
  - Use a similarity measure more appropiate to probability distributions.
  - Use a 2D histogram to show the relation betweeen topics instead of a scatter plot.
  - Find better statistics to show instead of the plot of a topic against the others.
- Change the topic modelling paradigm:
  - Change to non-parametric topic models.
  - Change to hierarchichal topic models.

Furthermore, we could consider the MeSH qualifiers when making our analysis to improve the correlation results obtained.

Chapter 3

Multiwords

We enhanced our model by adding the capability to handle *multiword expressions* or *phrases*.

3.1 Options

There is three standard techniques to implement multiwords:

1. Preprocess the corpus to find multiwords and feed the new corpus to an standard topic model.
2. Use a model that considers word dependency information to learn the topics, such as BTM [19], LDACOL [8], TNG [20] or TWC [10].
3. Post-process the output of an standard topic model to find multiwords appearing in the topics, such as with "turbo topics" [1].

We chose to implement the first option for various reasons:

- Given the size of our data, we cannot afford the complexity overhead incurred by topic models that use phrases as part of its framework.
- Processing the output of an standard model makes topics more readable but does not change the generated model.
- Preprocessing the input gives us the flexibility to migrate to new software and topic models using the same preprocessing chain.
- Once implemented in Bluima, multiword features can be easily turned on or off at will.

3.2 Analysis of the literature

A thorough analysis of the effect of multiword expressions in topic models was done in [11]. The results concerning our project are the following:

Unigrams vs. multiwords: Topic coherence is consistently improved when using multiwords.

Supplementation vs. Replacement: A replacement approach (where each occurrence of a multiword is replaced by its multiword token) yields better results than a supplementation approach (where each occurence of a multiword is replaced by its multiword token *and* its unigrams tokens).

Automatically generated vs. gold standard: Gold standard multiwords (those handpicked by an expert or obtained from a dictionary) outperform automatically generated multiwords (obtained by frecuency counts on the corpus).

Generality vs Specificity: The specifity of topics increases when increasing the number of multiwords.

Number of multiwords: Best results were obtained when using a moderate number of automatically generated multiwords (between 1K-10K).

Frequence of multiwords in the corpus: Replacing the top-1K automatically generated bigrams results in 2-5% of the new corpus being bigram tokens. This number raises to 12-16% when replacing the top-100K bigrams.

Frequence of multiwords in the output: On average, less than 10% of the top-10 topic terms were multiwords.

3.3 Implementation

For this part of the project we worked with documents composed by the title and abstract of articles obtained from the PubMed database. Considering the size and the vast amount of "real" topics existing in our corpus we believe using multiwords will help improve our results. We have chosen to use gold standard multiwords with a replacement approach. We use named entities obtained from different ontologies as multiwords to be identified in the corpus. The corpus contains 1 000 000 documents closely related to Neuroscience; every article has at least one MeSH descriptor under the category "Nervous System".

3.3.1 Bluima

The corpus is preprocessed using Bluima [18], a software based on the Apache UIMA$^{\text{TM}}$ framework. We offer some basic description of the relevant UIMA components for our project:

- *jCas:* An object designed to enclose a single document in the UIMA framework. The preprocessing is done by adding Annotations (defined below)

to the jCas of a document. The text of a document remains unchanged throughout the whole preprocessing.

- *Annotation:* An object that represents a single feature of the document. Annotations offer the possibility to set the indices indicating the beginning and end of the text spanned by the annotation. It can be subclassed to generate more specific Annotations. For example, we use the annotation Protein that apart from signaling where the annotation starts and ends has properties indicating the dictionary it came from, the canonical name of the protein, etc.

- *CollectionReader* An object that reads documents from the source (a simple text file in our case) and returns the jCas. Documents are read and processed one by one.

- *AnalysisEngine:* The primary processing component. Receives a jCas, analyzes its contents (document text and annotations) and performs some preprocessing. It could add new annotations, remove existing annotations or compute some statistic on the document. For example, we use the engine PunctuationAnnotator that reads through a document and annotates the ocurrences of a punctuation mark.

- *Pipeline:* A Bluima feature that allows to easily specify the processing steps to execute in UIMA. Essentially, it receives a text file with the name of the collection reader and the engines to execute and takes care of the initialization and execution. See an example in Appendix A

3.3.2 Preprocessing chain

The preprocessing is executed document by document in different stages; we offer a complete description of the preprocessing the corpus went through:[1]

1. Reading

   - OneDocPerLineReader2: Reads documents line by line from a tab separated text file in the format <PMID>\t<TITLE>\t<TEXT>. Creates the jCas by joining the text from <TITLE> and <TEXT>. Annotates the jCas with a Header annotation containing the title and the id of the document.

2. Standard preprocessing

   - OpenNLPHelper.getSentenceSplitter(): Biomedical oriented sentence splitting tool based on OpenNLP's MaxEnt SentenceDetector. Adds annotation Sentence.
   - OpenNLPHelper.getTokenizer(): Iterates over Sentences and invokes an standard tokenizer provided by Apache OpenNLP. Adds annotation Token.

[1] The real pipeline used is provided in Appendix A

- OpenNLPHelper.getPOSTagger(): Iterates over Sentences and invokes the Part-of-speech tagger provided by Apache OpenNLP. POS is added as a property of Token.
- BioLemmatizer: Biomedical oriented lemmatization tool. Adds the lemma as a property of Token using its POS.

3. Annotation retrieval

   - MeasureRegexAnnotators.getAllAED(): Regex-based measure extraction tool. Adds annotation Measure.
   - PruneMeasuresAnnotator: Prunes overlapping Measure annotations by removing one of them from jCas.
   - PunctuationAnnotator: Iterates over Tokens and adds annotation Punctuation to tokens which are only a punctuation mark.
   - SkipSomePosAnnotator: Iterates over Tokens and adds annotation POSSkip to tokens with uninteresting POS. Also adds POSVerb and POSAdverb to tokens that have a POS verb and adverb respectively.

4. Named entity recoginition

   - getConceptMapper("<DICTIONARY>"): Uses a lexical based named entity recognizer provided by Apache UIMA to match the words in DICTIONARY. Ignores capitalization and matches longest word possible. Adds an specific annotation depending on the dictionary with information about the matched term. The complete list of lexica used can be seen in Appendix A.

5. Choose Annotations to keep

   - ViterbiFilterAnnotator: Iterates over Sentences and encloses the Annotations in the shortest path with a Keep annotation. Decides between two Annotations in the shortest path based on confidence.

6. Normalization

   - BioLemmatizerNormalizerAnnotator: Iterates over Keeps and sets its normalized text to the lemma generated by the BioLemmatizer (normalizedText is a property of Keep).
   - MeasureNormalizerAnnotator: Sets the normalized text of Keeps enclosing a Measure annotation to "MEASURE_"+unit, where "MEASURE_" replaces any numerical value in the Measure. Removes Keeps that enclose a Measure without unit.
   - EntityNormalizerAnnotator: Sets the normalized text of Keeps that enclose a named entity to the canonical version of the term.

7. Annotation filtering

- PunctuationFilterAnnotator: Removes Keeps that enclose a Punctuation annotation.
- StopwordFilterAnnotator: Removes Keeps whose covered text matches a term in a small list of stopwords.
- AnnotationFilterAnnotator: Removes Keeps that enclose a POSSkip annotation.

8. Frequency counts
   - FrequencyFilterWriter: Counts the number of ocurrences of each Keep on the entire corpus (using its normalizedText). At the end writes the frequencies to a text file.
9. Writing
   - LdaCWriter: Writes the Keeps of each jCas to a text file in DCA format.
10. Frecuency filtering is done outside UIMA using the outputs from LdaCWriter and FrecuencyFilterWriter. Terms appearing less than 20 times were removed.

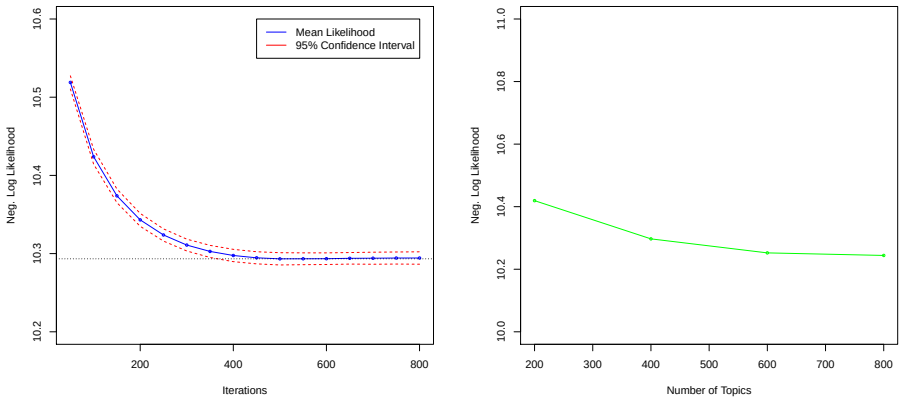
An extense list of collection readers and analysis engines have already been developed and are available in Bluima besides those used for this project.

3.4 Topic and iteration fitting

We performed topic and iteration fitting on this corpus generating the results in Figure 3.1. Iteration fitting was done using 400 topics in 10 fold crossvalidation with held-out data. Due to time constraints topic fitting was performed in the corpus with a 10% test size, i.e., crossvalidation was not used. As can be seen in the figures topic fitting needs to be run for a bigger number of topics and preferably using crossvalidation to reach a conclusive result.
We learned that a number of iterations greater than 450 is enough to assure the convergence of our model. On the other hand, we do not have convicing results yet for the number of topics; we expect the curve to keep descending as the number of topics augments.

3.5 Results

We trained two different topic models (600 topics, 600 iterations) on our 1M corpus: one using the corpus with minimal preprocessing as described in Section 2.1 and one using the corpus with the full preprocessing as described above. We analyze the results of the topic model generated by choosing an article at random and displaying its topic distribution as was performed in Section 2.2. In fact we will choose the same article analyzed in Section 2.2.



(a) Iteration fitting using 400 topics (b) Topic fitting using 700 iterations

Figure 3.1: Iteration and topic fitting results using 10-fold cross validation on 1M corpus.

3.5.1 Analysis of Article 13130504

For convenience, we present the abstract again in Figure 3.2. The article talks about the role of protein kinase B (Akt1) in the maintenance of cellular integrity during neuronal injuries. It is appropiate for our purpose because it uses specific technical terms that may have been recognized as named entities during our preprocessing.

Topic 385	**0.169**	**Topic 271**	**0.149**	**Topic 254**	**0.097**	**Topic 33**	**0.078**	**Topic 439**	**0.071**
kinase	0.115	cell	0.131	membrane	0.240	mitochondrial	0.124	microglia	0.101
phosphorylation	0.060	apoptosis	0.112	lipid	0.061	mitochondrion	0.062	microglial	0.066
protein	0.053	death	0.107	plasma	0.032	membrane	0.031	cell	0.051
activation	0.035	apoptotic	0.041	potential	0.016	potential	0.028	macrophage	0.049
pathway	0.028	induce	0.031	phospholipid	0.014	apoptosis	0.027	activation	0.040
inhibitor	0.027	caspase	0.022	raft	0.014	cell	0.027	activate	0.029
signaling	0.026	caspase-3	0.022	bilayer	0.013	cytochrome	0.020	inflammatory	0.018
tyrosine	0.022	bcl-2	0.016	protein	0.012	induce	0.020	expression	0.015
erk	0.019	activation	0.013	cholesterol	0.009	release	0.015	brain	0.014
mapk	0.018	dna	0.010	surface	0.008	permeability	0.010	immune	0.013

Table 3.1: First 5 topics related to article PMID 13130504 ordered by $p(z|d)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

We start by presenting the results for the corpus with minimal preprocessing in Table 3.1.[2] The document is related with 30 topics out of which the 5 topics presented are major topics $p(z|d) > 0.05$ and 9 topics have negligible probability ($p(z|d) < 0.01$). This is not a very good indicator given that it shows a mostly uniform topic distribution. The major topics appear to have high semantic

[2]Only the major topics are presented.

Akt1 protects against inflammatory microglial activation through maintenance of membrane asymmetry and modulation of cysteine protease activity.

In several cell systems, protein kinase B (Akt1) can promote cell growth and development, but the "antiapoptotic" pathways of this kinase that may offer protection against cellular inflammatory demise have not been defined. Given that early cellular membrane phosphatidylserine exposure is a critical component of apoptosis, we investigated the role of Akt1 during neuronal apoptotic injury. By employing differentiated SH-SY5Y neuronal cells that overexpress a constitutively active form of Akt1 (myristoylated Akt1), free radical-induced cell injury was assessed through trypan blue dye exclusion, DNA fragmentation, membrane phosphatidylserine exposure, protein kinase B phosphorylation, cysteine protease activity, and mitochondrial membrane potential. Membrane phosphatidylserine exposure was both necessary and sufficient for microglial activation, insofar as cotreatment with an antiphosphatidylserine receptor-neutralizing antibody could prevent microglial activity following neuronal loss of membrane asymmetry. Furthermore, expression of myristoylated Akt1 not only prevented cell injury through the prevention of membrane phosphatidylserine exposure and genomic DNA fragmentation but also inhibited microglial activation and proliferation that required the inhibition of caspase 9-, caspase 3-, and caspase 1-like activities linked to cytochrome c release. Interestingly, Akt1 modulation of membrane phosphatidylserine exposure was primarily through caspase 1 activity. Removal of Akt1 activity abolished neuronal protection, suggesting that Akt1 functions as a critical pathway for the maintenance of cellular integrity and the prevention of phagocytic cellular removal during neurodegenerative insults.

Figure 3.2: Abstract for article PMID 13130504.

relation with the article but do not show a great improvement when compared with the topics generated with the smaller corpus (see Table 2.2).
Similarly, we present the results for the corpus with multiword preprocessing in Table 3.2. This document is related with 17 different topics out of which the 6 presented are major topics ($p(z|d) > 0.05$) and 5 have a negligible probability ($p(z|d) < 0.01$). This is a good topic distribution given that our document is highly related to only a very small subset of the 600 topics generated. In addition to that, we also notice that in this case we have topics equivalent to those in the minimum preprocessing case but with more specialized words and better coherence.
As expected, adding multiword handling in the preprocessing of the corpus generates similar or better results than using minimal preprocessing. Similarly, we notice that results when using multiwords have improved when compared to those obtained with an smaller corpus.

Topic 12	**0.329**	**Topic 340**	**0.201**	**Topic 430**	**0.114**
apoptosis	0.090	Kinase	0.048	NEURON	0.078
CELL	0.076	activation	0.044	death	0.047
death	0.051	pathway	0.040	CELL	0.045
apoptotic	0.031	signal	0.035	neuroprotective	0.031
induce	0.024	phosphorylation	0.034	effect	0.030
caspase	0.020	mitogen-activated_protein_kinase_1	0.023	culture	0.025
apoptosis_regulator_Bcl-2	0.018	ERK	0.022	protect	0.023
activation	0.017	PROTEIN	0.021	induce	0.018
caspase-3	0.017	RAC-alpha_serine/threonine-protein_kinase	0.020	toxic_encephalopathy	0.018
deoxyuridine_5'-triphosphate_nucleotidohydrolase	0.011	inhibitor	0.018	toxicity	0.017
Topic 500	**0.074**	**Topic 351**	**0.054**	**Topic 573**	**0.054**
microglial_cell	0.070	role	0.168	cholesterol	0.085
macrophage	0.038	play	0.079	MEMBRANE	0.059
microglial	0.033	important	0.035	Fatty_acid	0.046
activation	0.030	function	0.033	lipid	0.027
Chemokine	0.021	critical	0.030	Lipoprotein	0.026
expression	0.021	regulation	0.026	raft	0.017
MICROGLIAL_CELL	0.020	suggest	0.024	ceramide	0.013
CELL	0.019	study	0.021	Phospholipid	0.012
inflammatory	0.014	evidence	0.017	periodontal_disease	0.011
MEASURE_GW	0.014	key	0.016	bilayer	0.011

Table 3.2: First 6 topics related to article PMID 13130504 ordered by $p(z|d)$ (bold). For each topic the list of the ten most probable words along with $p(w|z)$ is presented.

Chapter 4

Conclusion

4.1 Conclusion

We have succesfully applied standard topic models to corpora tightly related to Neuroscience. We have shown that even a small corpus (100 000 documents) can be used to train a topic model with moderately good results. We have used the MeSH descriptors assigned to PubMed articles to calculate a correlation measure between topics and MeSH descriptors with promising results. We used the information from the MeSH structure to calculate the correlation between so called MeSH graph nodes and topics producing similar results to those obtained when correlating with MeSH descriptors; although these results were not conclusive and further analysis may be required. Furthermore, interested in trying to desambiguate the appearance of a MeSH descriptor in different places of the MeSH hierarchy we attempted to correlate topics and MeSH tree numbers but results were poor due mostly to the construction of the MeSH structure. Lastly, we have implemented multiword handling for our preprocessing chain obtaining encouraging results even though due to time constraints a deeper analysis was not possible.

4.2 Future Work

Future work could be focused on various different directions. We list some of the more promising routes of action that could be taken.
We could use the preprocessed corpus of 1 million PubMed abstracts to calculate the MeSH correlation with descriptors and grpah nodes attempting to obtain definitve results on which of the two options offers an advantage. Also, results obtained from the MeSH correlations could be used to create a classification system that receives a document and returns a list of possible MeSH descriptors, this is not specially hard once a reliable topic model is generated. This classification system could also be used as a grading system for different correlation matrices to avoid the subjectivity of qualitative analysis and address

objectively which correlation results are better than others. An update to the topic model and software used could also be a good time investment. Some fine tuning in the preprocessing of documents can always be considered but at this point it may not be the most beneficial option.

4.3 Acknowledgements

I am thankful to both my supervisors Jean-Cédric Chappelier and Renaud Richardet who keep me on track and were always readily available to give me insights in topics ranging from high level concepts to grammar errors and without whom these work could not have been realized.

Appendix A

Pipeline

Bluima pipeline used to preprocess the 1M corpus.

```
#Annotation retrieval:
ae_java: ch.epfl.bbp.uima.ae.MeasureRegexAnnotators.getAllAED()
ae: ch.epfl.bbp.uima.ae.PruneMeasuresAnnotator
ae: ch.epfl.bbp.uima.ae.PunctuationAnnotator
ae: ch.epfl.bbp.uima.ae.SkipSomePosAnnotator
#NERs
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/age")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/disease")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/ionchannel")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/method")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/molecule")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/organism")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/region")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("blueonto1/sex")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("/brainregions/neuronames")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("/onto_baseline/cell")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("/onto_baseline/disease")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("/onto_baseline/protein")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("/onto_baseline/verb")
ae_java: ch.epfl.bbp.uima.LexicaHelper.getConceptMapper("/nif/nif")

#Choose annotations to keep (longest match rule)
ae: ch.epfl.bbp.uima.ae.ViterbiFilterAnnotator

#Normalization
ae: ch.epfl.bbp.uima.ae.BioLemmatizerNormalizerAnnotator
ae: ch.epfl.bbp.uima.filter.MeasureNormalizerAnnotator
 removeSimpleMeasure__java: true
ae: ch.epfl.bbp.uima.ae.EntityNormalizerAnnotator
```

```
#Annotation filtering
ae: ch.epfl.bbp.uima.filter.PunctuationFilterAnnotator
ae: ch.epfl.bbp.uima.filter.StopwordFilterAnnotator
ae: ch.epfl.bbp.uima.filter.AnnotationFilterAnnotator
 annotationClasses__java: new String[]{"ch.epfl.bbp.uima.types.POSSkip"}

#Frecuency counts
ae: ch.epfl.bbp.uima.filter.FrequencyFilterWriter
 outputFile: /home/michael/Documents/MasterProject/lda_mesh/corpora/1m_ns/1m_ns.vocab.raw

# Writing
ae: ch.epfl.bbp.uima.ae.output.LdaCWriter
 dcaFormat__java: true
 outputFile: /home/michael/Documents/MasterProject/lda_mesh/corpora/1m_ns/1m_ns.dca_corpus
 vocabularyOutputFile: /home/michael/Documents/MasterProject/lda_mesh/corpora/1m_ns/1m_ns.dc
 idsOutputFile: /home/michael/Documents/MasterProject/lda_mesh/corpora/1m_ns/1m_ns.pmids

ae: StatsAnnotatorPlus
 printEvery__java: 100
ae: GarbageCollectorAnnotator

# Only recommended for testing.
#ae: ch.epfl.bbp.uima.ae.output.BartWriter
# debug__java: true
# outputDir: /home/michael/Documents/MasterProject/lda_mesh/BluimaTest/bart

#ae: ch.epfl.bbp.uima.filter.KeepsDumper
# printCoveredText__java: true
```

Appendix B

Report Draft

Before arriving to the correlation functions used in Chapter 2 we used the formulas in this appendix. This is a working version of the results obtained with these formulas and is left here in case it could be useful in the future. It may be incomplete or incorrect.

B.1 MeSH Correlation

The probability of co-ocurrence of a MeSH descriptor and a topic on the corpus is given by:

$$p(m,z) = \sum_{d \in \mathcal{C}} p(z|m,d)p(m|d)p(d)$$

where d is a document, $\mathcal{C}$ is the corpus, z is a topic and m is a MeSH descriptor. Under the assumption that $p(m|d)$ is uniform on a given document[1], i.e., $p(m|d) = \frac{\mathbf{1}_{M(d)}(m)}{|M(d)|}$, that $p(z|m,d) = p(z|d)$ and that $p(d) = \frac{1}{|\mathcal{C}|}$ we obtain:

$$p(m,z) = \frac{1}{|\mathcal{C}|} \sum_{d \in \mathcal{C}} \frac{p(z|d)\mathbf{1}_{M(d)}(m)}{|M(d)|} \tag{B.1}$$

where $M(d)$ is the set of MeSH descriptors for document d and $\mathbf{1}_{M(d)}$ is the indicator function of $M(d)$.
When going through the summation on Equation B.1 we can choose to consider all the descriptors of a document or only the major ones, thus effectively assigning $M(d)$ to be the set of major descriptors. Likewise, we can choose to consider all the topics of a document or only the major ones, effectively assigning $p(z|d) = 0$ to non major topics. Major descriptors and major topics

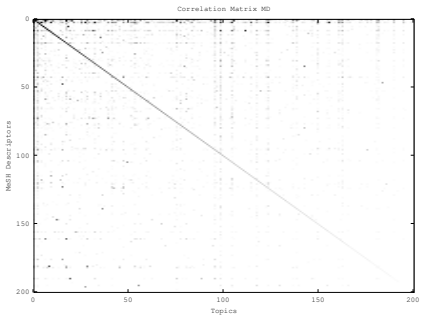
[1] This is certainly not true. For instance, $p(m|d)$ for major descriptors will presumably be higher than for not majors. However, as we have no further information about the tagging process we make this general assumption.

(abbreviated MD and MT) were described in Section 2.2.
We use this probability to generate a so called *correlation matrix* between topics and MeSH descriptors[2].

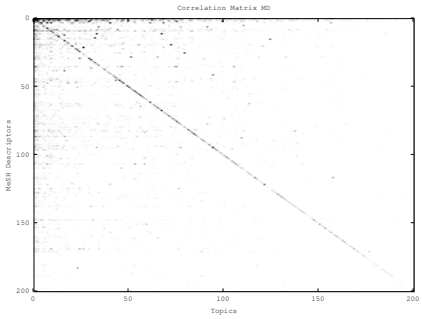
B.1.1 Visualization

Rows (MeSH descriptors) and columns (topics) on the correlation matrix could be reordered to reveal hidden structure in the data[3]. We enlist the different methods considered in this project along with an example:

- Maximize the values on the diagonal of the matrix and crop the number of rows to match the number of topics (200 in this case).

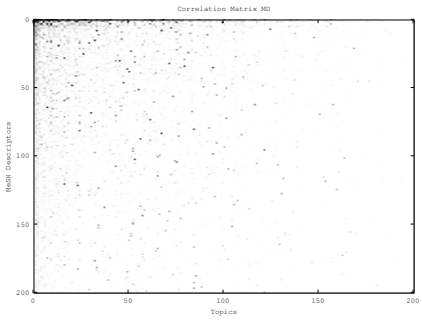


- Order columns on descending order (value of a column is the sum over its rows). Once columns are fixed reorder rows such as to maximize the diagonal values. Crop the number of rows to match the number of topics (200 in this case).

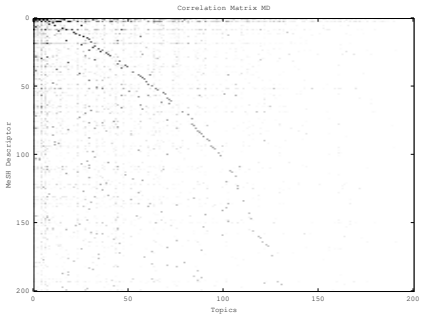


- Order columns on descending order (value of a column is the sum over all its rows). Once columns are fixed reorder rows in a similar fashion. Crop the number of rows to match the number of topics (200 in this case).

[2]The term correlation is used throughout this section to refer to $p(m, z)$.
[3]This is a research problem in itself generally referred as *seriation* or *sequencing*.



- Push high values on the matrix to the upper left corner such that the highest kth value is at worst on a k-sized box on the upper left corner. Crop the number of rows to match the number of topics (200 in this case).



- Use the R seriation package [9] with its different heuristics to reorder the matrix. For these examples we used the matrix on B.1a as input because the package was not able to handle the entire matrix (size 11642 x 200).

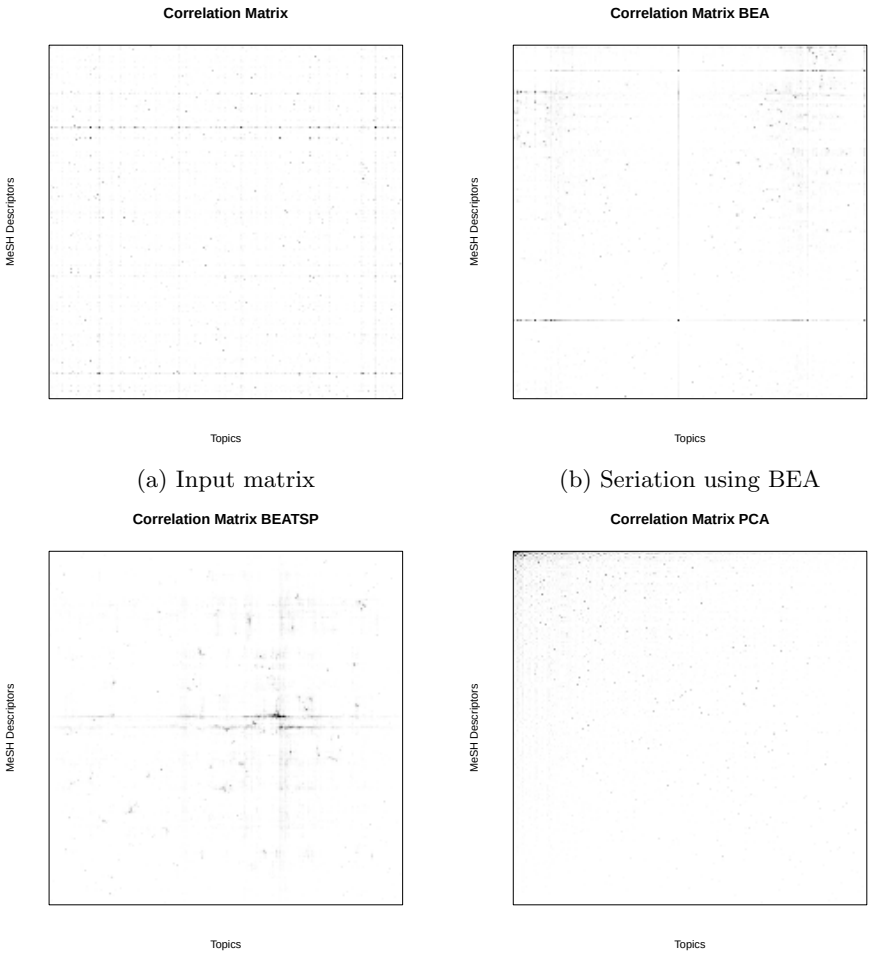
We chose the first option to present our results given that it is simple, fast and good enough to show the relation between topics and descriptors.

B.1.2 Results

We generated four different correlation matrices: one using all descriptors and all topics, one using all descriptors but only major topics, one using only major descriptors and all topics and one using only major descriptors and only major topics.

Results can be found on Figure B.2. Every matrix has been reordered such as to maximize the value in the diagonal and cropped to fit the number of topics, in this case 200.

We observe a clear line in the diagonal of our matrices indicating that there are various topic-descriptor pairs with a high correlation value. Certain general topics such as topic 126 shown in Table 2.1 have a high correlation with the majority of descriptors; this fact is illustrated in the matrices by dark vertical lines. Likewise, general descriptors such as "Brain" relate to a big proportion of topics, which is shown in our matrices by dark horizontal lines.



(a) Input matrix

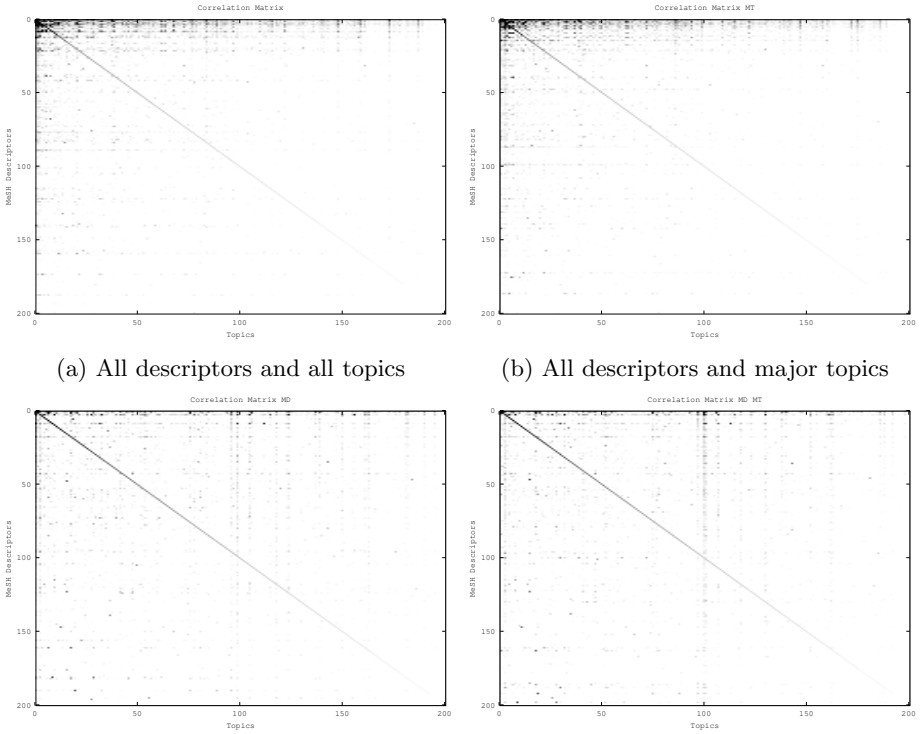
(b) Seriation using BEA

(c) Seriation using BEA_TSP

(d) Seriation using PCA

Figure B.1: Reordering using seriation

We are interested on matrices where descriptors relate strongly to only a few number of topics, i.e, a matrix with high values on the diagonal and few other spots but low values elsewhere. This kind of matrix allows to identify each MeSH descriptor by a probability distribution over a small set of topics.
Matrices MD and MD MT appear to show clearer results, pointing towards the observation that selecting only major descriptors to compute this type of correlation yields better results than selecting all of them. This is an artifact of the assumptions made when deriving Equation 2.1, it seems that our uniformity assumption on $p(m|d)$ is more accurate when considering only the major descriptors for the document. Altogether we believe the best outcome is obtained



(a) All descriptors and all topics

(b) All descriptors and major topics

(c) Major descriptors and all topics

(d) Major topics and major descriptors

Figure B.2: Correlation matrices for the four different descriptor and topics combinations used to calculate $p(m, z)$ on the 100K corpus (as explained on Section subsec:100KResults). Reordered as to maximize the diagonal values and cropped to match the number of topics (200)

on matrix MD and proceed to analyze some of its features on the next section.

Analysis of Correlation Matrix MD

We enlist different descriptor-topic pairs with their correlation values to verify that our results are semantically valid. Table B.1 shows the first values of the diagonal of correlation matrix MD. Probability values are calculated with Equation B.1.

We observe that the highest correlation value in the entire matrix associates the most common descriptor ("Brain") and the most common topic (126) in the corpus even though they are not conceptually related; this is an effect of the general topics produced by the standard topic model and is not dependent on the selected corpus. In general, common descriptors and topics have higher correlation values than uncommon ones, which does not imply that we will consistently produce bad results when dealing with common descriptors and topics;

MeSH Descriptor	Topic	p(m, z)
(001921) Brain	(126) study, result, effect, show, suggest, human, change, present, increase, previous	173.13
(008279) Magnetic Resonance Imaging	(80) imaging, mr, image, mri, magnetic, signal, resonance, sequence, brain, contrast	135.84
(009474) Neurons	(119) role, function, mechanism, system, pathway, play, neuronal, involve, important, regulate	118.76
(014793) Visual Cortex	(185) visual, motion, field, cortex, stimulus, area, orientation, spatial, receptive, v1	83.743
(017209) Apoptosis	(60) cell, death, apoptosis, neuronal, neuron, induce, apoptotic, activation, dna, culture	82.875

Table B.1: Highest correlated topic-descriptor pairs on the diagonal of matrix MD for 100K corpus. Probabilities are not normalized with respect to $|\mathcal{C}|$.

in fact, as can be seen in the other rows of Table B.1, results are quite good.
As long as the topic model generates a small number of "general purpose" topics and all descriptors appear relatively often in the corpus results will be positive. Nonetheless, we will not be able to completely remove this effect given that our corpus will certainly have common descriptors and our model will certainly produce general topics.
As a further experiment we show the 100th, 500th, 2 000th and 10 000th highest value in the whole correlation matrix (11642 descriptors x 200 topics) on Table B.2.

MeSH Descriptor	Topic	p(m, z)
(001931) Brain Mapping	(140) method, datum, image, analysis, brain, technique, estimate, model, approach, result	37.129
(010902) Pituitary Gland	(172) gnrh, lh, hormone, pituitary, secretion, fsh, reproductive, release, gonadotropin, female	14.769
(005625) Frontal Lobe	(144) disease, 's, ad, alzheimer, tau, dementia, patient, alpha-synuclein, body, brain	5.6611
(018698) Glutamic Acid	(183) MEASURE_, al., j., respective, study, micro, previous, hr, type, find	1.5397

Table B.2: 100th, 500th, 2 000th and 10 000th highest value of entire matrix MD for 100K corpus. Probabilities are not normalized with respect to $|\mathcal{C}|$.

As expected, MeSH descriptors appearing on this table are not as general as before but still have high correlation values with the given topic. As the correlation value decreases the semantic correlation also declines; as a matter of fact the last two descriptor-topic pairs do not appear to make much sense although we have to point out that these descriptors probably have higher correlations with topics other than the one listed here. Topic 183 is again a so called "general purpose" topic and the correlation with descriptor "Glutamic Acid" is not very meaningful.
As a whole, these results are good and show a clear semantic correlation between topics and descriptors, which is remarkable given that we did not perform any special tuning on the corpus or the model. We expect these results to get better as we increase the size of the corpus and refine the model.
It is also worth noticing that we obtained better results than previously pre-

sented [21].

B.1.3 Applications

The generated topic model and correlation matrix could be used for various practical tasks; we enlist some of them:

- Descriptor prediction for a document d could be performed by building a ranking system using $p(m|d)$.
- Major descriptor prediction for a document d with a given set of descriptors $M(d)$ could be done in a similar fashion.
- New relations on the MeSH hierarchy could be retrieved using information measures such as the symmetric KL divergence between descriptors.

Some of this experiments were performed on a smaller corpus in [16].

B.2 MeSH Correlation with graph nodes

In this section we use our knowledge of the MeSH structure to compute $p(n, z)$, the probability of co-ocurrence of graph node n and topic z. Nodes in the structure are identified by a single descriptor and may have various tree numbers as explained on Section 1.2.2.

(NOTE: Choose an option, delete the section title and leave the text. Option 1 is the one we have discussed)

B.2.1 Option 1

The probability of co-ocurrence of a graph node and a topic is given by:

$$p(n, z) = \sum_{d \in \mathcal{C}} \sum_{m \in M(d)} p(z|n, m, d)p(n|m, d)p(m|d)p(d)$$

Assuming that:

- $p(m|d)$ is uniformly distributed.
- $p(z|d, m, n) = p(z|d)$. This holds when $M(d)$ contains MeSH descriptor m and $N(m)$ contains node n, which is always true on our formula.
- $p(n|m, d) = p(n|m)$
- $p(n|m) = \mathbf{1}_{N(m)}(n)$

We obtain:

$$p(n, z) = \frac{1}{|\mathcal{C}|} \sum_{d \in \mathcal{C}} \sum_{m \in M(d)} \frac{p(z|d)\mathbf{1}_{N(m)}(n)}{|M(d)|} \tag{B.2}$$

B.2.2 Option 2

The probability of co-ocurrence of a graph node and a topic is given by:

$$p(n,z) = \sum_{d \in \mathcal{C}} \sum_{m \in M} p(z|n,m,d)p(n|m,d)p(m|d)p(d)$$

Assuming that:

- $p(m|d)$ is uniformly distributed.
- $p(z|d,m,n) = p(z|d)$. This holds when $M(d)$ contains MeSH descriptor m and $N(m)$ contains node n, which is always true on our formula.
- $p(n|m,d) = p(n|m)$
- $p(n|m)$ is uniformly distributed.

We obtain:

$$p(n,z) = \frac{1}{|\mathcal{C}|} \sum_{d \in \mathcal{C}} \sum_{m \in M(d)} \frac{p(z|d)\mathbf{1}_{N(m)}(n)}{|M(d)||N(m)|} \tag{B.3}$$

B.2.3 Option 3

The probability of co-ocurrence of a graph node and a topic is given by:

$$p(n,z) = \sum_{d \in \mathcal{C}} p(z|n,d)p(n|d)p(d)$$

We define $p(n|d)$ as:

$$p(n|d) = \frac{1}{\sum_{m \in M} |N(m)|} \sum_{m \in M(d)} \mathbf{1}_{N(m)}(n) \tag{B.4}$$

Assuming that $p(z|d,n) = p(z|d)$, we derive:

$$p(n,z) = \frac{1}{|\mathcal{C}|} \sum_{d \in \mathcal{C}} \sum_{m \in M(d)} \frac{p(z|d)\mathbf{1}_{N(m)}(n)}{\sum_{m \in M(d)} |N(m)|} \tag{B.5}$$

(NOTE: End of options. From here on is the same content for any option)
where n is a graph node, m is a MeSH descriptor, z is a topic, $\mathcal{C}$ is the corpus, $M(d)$ is the set of MeSH descriptors for document d, $N(m)$ is the set of nodes that appear in higher levels of the MeSH hierarchy and connect to the node identified by m (includig the node itself) and $\mathbf{1}_{N(m)}$ is the indicator function of $N(m)$.
The set $N(m)$ can be constructed using the tree numbers of the node identified by m; for instance the node identified by "Face" has tree number A01.456.505, therefore: N("Face") = {A01 "Body Regions", A01.456 "Head", A01.456.505

"Face"}. Applying the same procedure we could also obtain: N("Palate") = {"Stomatognathic System", "Mouth", "Palate", "Jaw"}[4].
We use this probability to construct correlation matrices between topics and nodes. As a node is uniquely identified by a MeSH descriptor our matrix will have rows representing MeSH descriptors and columns representing topics; similar to matrices in Section 2.3. Notice however that in this case MeSH descriptors that do not appear directly in the corpus could appear as a row in the matrix given that this descriptors may be in the set $N(m)$ for a descriptor m that does appear in the corpus. Therefore, we will potentially produce bigger matrices than when correlating topics with descriptors.

B.2.4 Results

As explained in Section B.1.2 we could create four different correlation matrices based on the data used in the computation of $p(n, z)$. In this section we only generated matrix MD, it is, we used all topics but only major descriptors. We decided to continue with this option because it showed the more promising results.
(NOTE: Yet to write after deciding which option is chosen above)

B.2.5 Correlation with tree numbers

We could also correlate topics directly to tree numbers, it is, compute $p(t, z)$ where t is a tree number and z is a topic. The formula is similar to B.5 and is derived in a similar fashion:

$$p(t, z) = \frac{1}{|\mathcal{C}|} \sum_{d \in \mathcal{C}} \sum_{m \in M(d)} \frac{p(z|d)\mathbf{1}_{T(m)}(t)}{|M(d)||T(m)|} \tag{B.6}$$

where t is a tree number, $T(m)$ is the set of tree numbers that compose a tree number of m (including the tree numbers of m) and other parameters are defined as before.
$T(m)$ can be calculated using the tree numbers assigned to the descriptor m. For example, for descriptor "Face" with tree number A01.456.505 N("Face") = {A01, A01.456, A01.456.505}. In the same way, T("Palate") = {A14, A14.549, A14.549.617, A14.521, A14.521.658}. Notice that not all tree numbers of descriptor "Mouth" appear in T("Palate") but only those (A14.549) that directly connect to a tree number on "Palate"[5].
Using Equation B.6 we could generate a correlation matrix having tree numbers as rows and topics as columns. This kind of correlation could help to dicriminate between different meanings of a descriptor; for example, one could argue that different tree numbers for the same MeSH descriptor refer to different concepts or have a particular focus and that this distinction will be captured by the relation between the different tree numbers and the topics. In practice, this

[4]Descriptors used in this examples appear in Figure 1.2.
[5]Descriptors used in this examples appear in Figure 1.2.

could allow to tag Pubmed articles directly with tree numbers or to suggest the creation of a new MeSH descriptor when a current descriptor has more than one concept attached to it.

Bibliography

[1] D. Blei and J. Lafferty. Visualizing topics with multi-word expressions. arXiv:0907.1013 [stat.ML], 2009.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Wray Buntine. DCA 0.202: Discrete component analysis software. Software documentation, Statistical Machine Learning Group, NICTA, Canberra, Australia, July 2009.

[4] Wray Buntine. DCA 0.202 user guide. User guide, Statistical Machine Learning Group, NICTA, Canberra, Australia, July 2009.

[5] Wray Buntine. Estimating likelihoods for topic models. In *Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning*, ACML '09, pages 51–64, Berlin, Heidelberg, 2009. Springer-Verlag.

[6] Wray Buntine. hca. User guide, Statistical Machine Learning Group, NICTA, Canberra, Australia, November 2013.

[7] Wray Buntine and Aleks Jakulin. Applying discrete pca in data analysis. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 59–66, Arlington, Virginia, United States, 2004. AUAI Press.

[8] Thomas L. Griffiths, Joshua B. Tenenbaum, and Mark Steyvers. Topics in semantic representation. *Psychological Review*, 114:211–244, 2007.

[9] Michael Hahsler, Kurt Hornik, and Christian Buchta. Getting things in order: An introduction to the r package seriation. *Journal of Statistical Software*, 25(3):1–34, March 2008.

[10] Wei Hu, Nobuyuki Shimizu, Hiroshi Nakagawa, and Huanye Sheng. Modeling chinese documents with topical word-character models. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 345–352, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[11] Jey Han Lau. *Improving the utility of topic models: an uncut gem does not sparkle*. PhD thesis, The University of Melbourne, Melbourne, Australia, June 2013.

[12] Haibin Liu, Tom Christiansen, William A. Baumgartner, and Karin Verspoor. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*, 3(1):3+, 2012.

[13] U.S. National Library of Medicine. Fact Sheet medical subject headings (MeSH). `https://www.nlm.nih.gov/pubs/factsheets/mesh.html`. Accessed: 2014-04-02.

[14] U.S. National Library of Medicine. NLM Medical Text Indexer. `http://ii.nlm.nih.gov/MTI/index.shtml`. Accessed: 2014-04-02.

[15] U.S. National Library of Medicine. Searching PubMed with MeSH. `http://ii.nlm.nih.gov/MTI/index.shtml`, Chicago, United States, November 2013. Accessed: 2014-04-02.

[16] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. Using topic models to interpret MEDLINE's medical subject headings. In Ann Nicholson and Xiaodong Li, editors, *AI 2009: Advances in Artificial Intelligence*, volume 5866 of *Lecture Notes in Computer Science*, pages 270–279. Springer Berlin Heidelberg, 2009.

[17] Project Management Institute. *A guide to the Project Management Body of Knowledge (PMBOK guide)*. Project Management Institute, fifth edition, 2013.

[18] Renaud Richardet, Jean-Cédric Chappelier, and Martin Telefont. Bluima: a uima-based nlp toolkit for neuroscience. In Peter Klügl, Richard Eckart de Castilho, and Katrin Tomanek, editors, *UIMA@GSCL*, volume 1038 of *CEUR Workshop Proceedings*, pages 34–41. CEUR-WS.org, 2013.

[19] Hanna M. Wallach. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 977–984, New York, NY, USA, 2006. ACM.

[20] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 697–702, Washington, DC, USA, 2007. IEEE Computer Society.

[21] Marc Zimmermann. UIMA integration of topic models. Semester project report, Artificial Intelligence Laboratory, EPFL, Lausanne, Switzerland, February 2013.