sciendo

Supporting secondary research in early drug discovery process through a Natural Language Processing based system

Alina POPA
The Bucharest University of Economic Studies, Bucharest, Romania
dragomir.alina.alexei@gmail.com

Abstract. *Last decades were characterised by a constant decline in the productivity of research and development activities of pharmaceutical companies. This is due to the fact that the drug discovery process contains an intrinsic risk that should be managed efficiently. Within this process, the early phase projects could be streamlined by doing more secondary research. These activities would involve the integration of chemical and biological knowledge from scientific literature in order to extract an overview and the evolution of a certain research area. This would then help refine the research and development operations.*
Considering the vast amount of pharmaceutical studies publications, it is not easy to identify the important information. For this task, a series of projects leveraged the advantages of the open pharmacological space through state-of-the-art technologies. The most popular are Knowledge Graphs methods. Although extremely useful, this technology requires increased investments of time and human resources. An alternative would be to develop a system that uses Natural Language Processing blocks. Still, there is no defined framework and reusable code template for the use-case of compounds development.
In this study, it is presented the design and development of a system that uses Dynamic Topic Modelling and Named Entity Recognition modules in order to extract meaningful information from a large volume of unstructured texts. Moreover, the dynamic character of the topic modelling technique allows to analyse the evolution of different subject areas over time. In order to validate the system, a collection of articles from the Pharmaceutical Research Journal was used.
Our results show that the system is able to identify the main research areas in the last 20 years, namely crystalline and amorphous systems, insulin resistance, paracellular permeability. Additionally, the evolution of the subjects is a highly valuable resource and should be used to get an in-depth understanding about the shifts that happened in a specific domain.
However, a limitation of this system is that it cannot detect association between two concepts or entities if they are not involved in the same document.

Keywords: Natural Language Processing, Dynamic Topic Modelling, Research Trends, Drug Discovery, Named Entity Recognition

Introduction

In one of the strategy studies from recent years made by KPMG, a big consulting firm, focused on the business model that is applicable to the pharmaceutical industry (Pharma), the consulting company states that the marketplace is undergoing huge changes (KPMG, 2017). Traditionally, big Pharma companies' business model was including the whole value chain starting from research and development (R&D) and finishing with licensing and commercialization. According to the same study, the traditional business model is currently experiencing a decline. This downturn was predicted and captured by a series of report and

articles from domain experts and consulting firms (Gilbert et al., 2003; PwC, 2009; Pammolli, 2011; Stott, 2017; Deloitte, 2020). One of the critical decisions a Pharma company should take in order to overcome this challenge is to improve their R&D productivity. It is no longer profitable for a player on pharmaceutical market to maintain the whole chain of bringing innovative compounds from laboratory to the marketplace (Philippidis, 2015). Still, in order not to fall in the trap of the so-called Innovators' dilemma (Alcantara et al., 2012; O'Reilly & Tushman, 2016), companies should change and improve their own processes and make restructuring decisions. The dilemma appears when one of the leaders of the market is questioning about the long-term turnover that innovations or new approaches could generate and, in that case, could cannibalize their own business.

In Pharma, the solution to the innovation crisis is not due to the lack of technology, but a new business model (Fleming, 2018). According to the same author, the main aspect of the R&D profitability decline is the intrinsic uncertainty and the outdated method of operating the drug discovery process. The main need for companies is to find a way of making the drug discovery more effective, leaner and faster by leveraging the existing knowledge efficiently (Rizzo et al., 2013). In the same time, pharma companies should keep a close watch at what is competition doing and what are the overall trends in a certain research area.

The total R&D investments are distributed quite uniformly throughout the development timeline with the difference that at earlier phases, there is a need of more and smaller investments in many projects than in the later phases, this due to the natural attrition within the R&D pipeline (Stott, 2017). As a response for reducing development risk in the early phase projects, Pharma companies are shifting focus from extensive internal R&D activities to open external research community by using the "open pharmacological space" (Vlijmen, 2016) through state-of-the-art technologies. This allows to run multiple programs in a more flexible way and at relatively lower cost, by implying a series of secondary research activities.

Using Machine Learning, Natural Language Processing and Knowledge Graphs (KG) to extract concepts that are later on linked together through semantic models, Pharma companies can easily search and relate areas like research topics, procedures, protein targets, biological processes etc. One of the popular methodologies refers to developing KG (Siebert, 2020). Examples of such projects are Bio2RDF (Belleau et al., 2008), Chem2Bio2RDF (Chen et al., 2010) and Open PHACTS (Groth et al., 2014). Although extremely useful and valuable, this type of systems requires a huge amount of maintenance and development resources. In order to build an integrated KG involves the creation of large semantic networks that link together data from multiple and different sources. This is a challenging task because of the different data formats and types as well as occasionally contradictory facts (Chen et al., 2014). Another limitation is that KG are not dynamical in nature and don't give an overview on the overall trend in a certain area. The second methodology of choice to facilitate knowledge extraction is using topic modelling techniques. This showed to be successful in a series of applications from statistical modelling of biomedical corpora (Blei et al., 2006), finding complex biological relationships (Wang et al., 2011), linking genetic basis of side-effects of drugs (He et al., 2011), evaluating disease similarity (Frick et al., 2015) to predicting adverse drug reaction (Xiao et al., 2017). Topic modelling holds advantages like being easy to understand and develop, bringing time efficiency and possibility of applying it in the areas with little or no prior knowledge (Wood et al., 2017). The fact that it represents a statistical method makes it reliable (King and Lowe, 2003) and objective (Mo et al., 2015), so that it

allows analysis replicability. Although all the stated advantages, there is no clearly defined nor coded framework that would allow using topic modelling. Same time, all of the previously studies were using the static version of the topic modelling approach.

The most significant source of knowledge is embedded in published literature (Chen et al., 2014). In order to be efficient in the early stages of the drug discovery process, Pharma companies need a tool that would allow the review, categorization and trend extraction from an extensive pharmacological corpus. In this paper, we try to design and develop a system that would answer two major objectives in the area of drug discovery through usage of NLP blocks: 1) what are the preeminent research areas in pharma and how did they evolve through time and 2) what are the organizations and centres that undergo drug discovery research in the main study areas.

The methodology used consists of a dynamic topic modelling (DTM) approach combined with an entity recognition engine. DTM will help understand the way a certain research area evolved in terms of procedures and protocols, protein targets, used compounds, discovered relationships while the organization entity extraction will map these efforts with market players that could be considered as competitors or potential research partners.

Literature review

Natural Language Processing (NLP) is a branch of artificial intelligence that makes it possible for computers to understand, process and generate language the same way people are able to do. NLP consists from a series of different and complementary tasks like automatic information summarization, translation, named entity recognition, relationship extraction, sentiment analysis, topic segmentation (Daelemans and Hoste, 2002).

Topic modelling

Topic modelling has proven to be an effective tool for exploratory analysis of a large number of papers in a reliable, fast and reproducible way (Asmussen and Moller, 2019). From a technical perspective, topic modelling refers to a group of unsupervised machine learning algorithms that infer the latent structure behind a collection of documents. This means that there is no prior knowledge about topics included in an article and the method cannot leverage information about correct answers. The intuition behind topic models is that each article consists of a series of topics. A topic then refers to a collection of words or terms specific to it.

Some examples of modelling approaches are: Latent Semantic Analysis (LSA) (Deerwester et al., 1988), Probabilistic Latent Semantic Indexing (pLSI) (Hoffman, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Hierarchical Topic Modelling (hLDA) (Griffiths et al., 2004), Dynamic Topic Modelling (DTM) (Blei and Lafferty, 2006), Supervised Topic Modelling (sLDA) (Mcauliffe and Blei, 2007). For the current study, the dynamic LDA modelling approach was selected, because it is answering the requirements of the expected results, namely extracting topics and their evolution in terms of keywords and terms.

Notation and terminology

Before specifying the mathematical model associated with the topic approach, we define the following terms:

- A *word w* is the basic unit of text data. A word is an element of a vocabulary and is represented as a one-hot encoded vector of length of vocabulary. This means that the vector will have a value of 1 at word's position and 0 elsewhere. As an example, imagine having the vocabulary consisting of words {marketing, communication, efficiency, promotions}, the word "communication" is then codified as [0,1,0,0].
- A *document d* represents a sequence of *N* words noted by d = $(w_1, w_2, w_3, ..., w_N)$.
- A *corpus* is a collection of *M* documents noted by D= $\{d_1, d_2, d_3, ..., d_M\}$, where d_m is the *m*-th document in the collection.

Dynamic Topic Modelling through dynamic Latent Dirichlet Allocation Method

In order to be able to use the DTM approach, it is needed first to review the underlying statistical assumptions of a static topic model like Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

The LDA model is a probabilistic generative method that extracts topics from a collection of papers. The LDA model assumes that a document contains a number of topics, so it is a probability distribution over topics. Each topic is defined as a probability distribution over words of a vocabulary. The method analyses the words in each paper and calculates the joint probability distribution between the observed elements (the words in the paper) and the unobserved ones (the hidden structure of the topics in the paper).

The *generative process* is described as follows (Blei, 2012):

- Primary *assumptions*:
  - There is a number of *K* topics in the documents collection
  - For each of the $k \in K$ topic, there is a distribution over words, $\beta_k \sim Dir(\eta)$
- *Document generation* process:
  - For each document $d = 1,..,M$, that is generated, is drawn a topic distribution that will be present in the document, $\theta_m \sim Dir(\alpha)$.
    - Based on this topic distribution, a topic is being randomly chosen $z_{m,n} \sim Multinomial(\theta_m)$
    - Based on the word distribution related to the chosen topic, a word is chosen randomly from the corresponding distribution over the vocabulary $w_{m,n} \sim Multinomial(\beta_k / k = z_{m,n})$
    - Previous two steps are iterated until the whole generation of the document is completed.

Thus, the main feature of LDA is the assumption that all documents in the collection have the same set of topics, but each document contains those topics in different proportions, from 0 to 1. In reality, we can only observe the words in a document, and we do not know the distributions that formed the basis of its generation (K, θ_m, β_k). Topics themselves, topics and words distributions are hidden variables. However, what is important to note is that the LDA process defines a joint probability distribution on both hidden variables (topics in a document) and observed variables (words in the document) as shown in figure 1. This joint distribution is used to deduce the hidden variables given the observed variables.

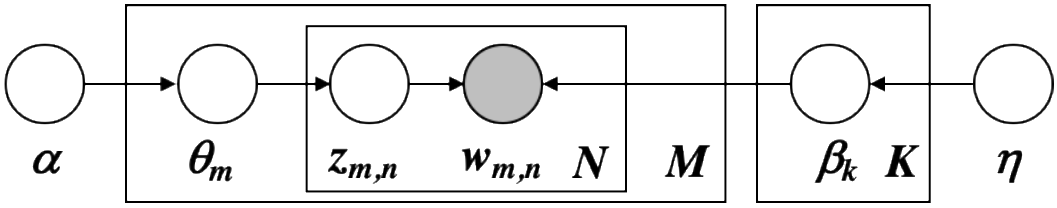

Figure 1. LDA Plate model notation

Source: Adaptation after figure 4 from Blei M. David. (2012), "Probabilistic Topic Models" Communications of the ACM, Vol. 55 Nr. 4, p. 81

From (Blei and Lafferty, 2006), it is shown that LDA is not applicable to sequential modelling in the form from *figure 1*, because the main Dirichlet components are not suitable for this task. As a response to this challenge, authors decide to chain multiple static LDA models and apply a random walk approach on the parameters.

This way, in DTM, the topic distribution over words for each topic β_k evolves with a Gaussian noise following the structure:

$$\beta_{t,k} | \beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I) \quad (1)$$

Same dynamic character is applied to the parameter of the Dirichlet distribution, α. As stated above, from $Dir(\alpha)$, distributions over topics that will be present in a certain document are drawn. In order to express uncertainty over these proportions, the sequential structure between different time models is captured with a simple dynamic model and it becomes a logistic normal distribution:

$$\alpha_t | \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I) \quad (2)$$

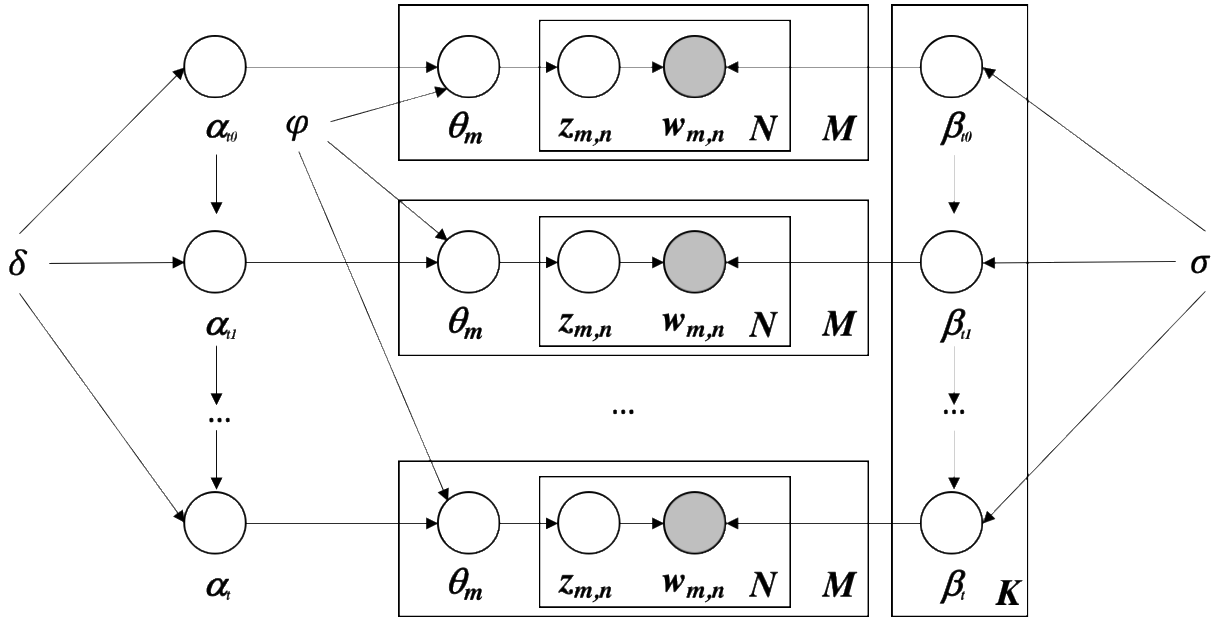

Figure 2. Graphical representation of a Dynamic Topic Model, where parameters of the Dirichlet distributions evolve over time

Source: Adaptation after Figure 1 from Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).

This way, through chaining together topics ($\beta_{t,k}$) and distribution of topic proportion (α_t), DTM represents a sequentially tied collection of topic models (*figure 2*) and the g*enerative process* for time slice *t* changes to:

- For each time slice *t:*
  - Draw topics*:* $\beta_t | \beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$
  - Draw topics' proportion: $\alpha_t | \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$
  - For each document:
    - Draw a topic distribution that will be present in the document $\theta \sim$ N (α_t, a^2I)
    - For each word:
      - Draw a topic $Z \sim Multinomial(\pi(\theta))$
      - Draw a word $w_{t,m,n} \sim Multinomial(\pi(\beta_t | k = Z))$, where π is a mapping under a soft-max transformation. This refers to the usual representation of a multinomial distribution by its mean parameterization.

Being a problem from the Bayesian paradigm, it is needed to compute the conditional distribution of the topic structure given the observed documents, or the so-called *posterior* distribution (Bhadury et al., 2016):

$$p(\alpha, \theta, \beta, Z | w, \sigma, \delta, \varphi) \propto \prod_{t=1}^{T} N(a_t | a_{t-1}, \delta^2 I) \times \prod_{k=1}^{K} N(\beta_{k,t} | \beta_{k,t-1}, \sigma^2 I) \times$$
$$\prod_{m=1}^{M} N(\theta_{t,m} | a_t, \varphi^2 I) \times \prod_{n=1}^{N_{t,m}} Mult(Z_{t,m,n} | \pi(\theta_{t,m})) Mult(w_{t,m,n} | \pi(\beta_{k,t} = Z_{t,m,n}) \quad (3)$$

As exact or analytical posterior inference is intractable due to the no conjugacy of the Gaussian and multinomial models, there is the need of using other methods to fit the model. Blei and Lafferty (2006) propose a variational Kalman filtering approach to infer the posterior distribution.

Methodology

System design

The aim of the paper is to create a specific framework and put in place a reusable system which can be applied in the context of an exploratory secondary research in the area of pharmacological studies. This should represent a tool that can be used in order to extract the research trends in a topic of interest, easily select the corresponding papers and get a situation of what are the organizations and companies that invest in the topic.

The method provided in this paper is a framework which reflects an end-to-end pipeline starting from getting the data to presenting the results. The framework contains established procedures for how to clean and process data, so that the main contribution is by putting in place together known techniques in a structured manner that can be reused later. The outcome of the system is an organized dataset where papers are assigned topics and topics are presented as distribution of terms over the time slices.

The developed system design is illustrated in figure 3. The corresponding code is located at: github[1]. The whole system was written in *python* language and the most important packages used in each of the steps are referred to in the text.

[1] https://github.com/alinkamalvinka/pharma_dtm

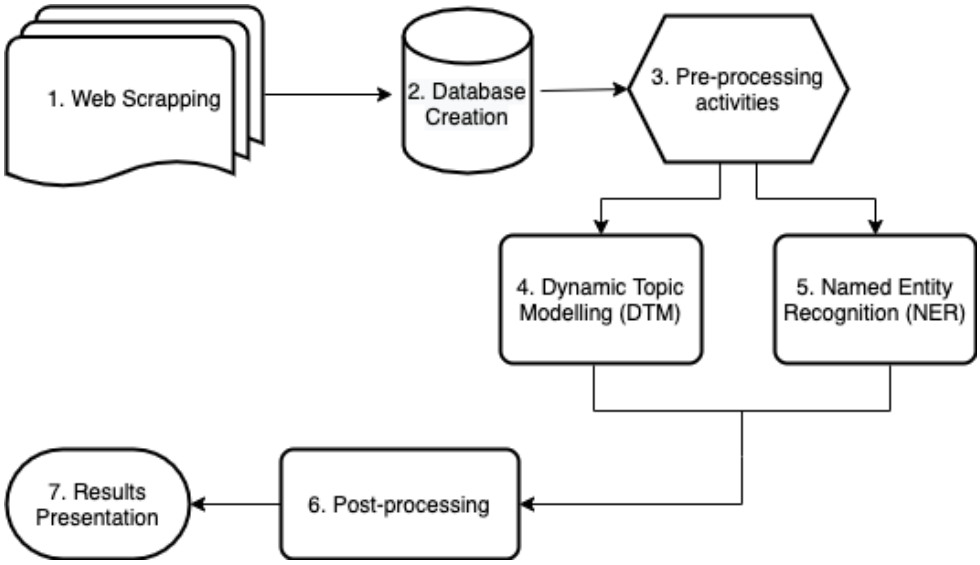

Figure 3. System design of the tool for extracting research area evolution and organizations and companies that are undergoing it.

Web scrapping

The code to scrap abstracts, titles and funding information from Springer portal is using *urllib*[2] and *PyQuery*[3] python packages. The code is organized in such a way that user needs to specify just the journal id from the Springer platform, the start year and the period over which he wants to extract data. This means that one is able to extract data from any other journal of interest that is present on the Springer platform. In the first step of creating the database, the articles are extracted in electronic format and their abstracts and titles are parsed. By parsing, it is identified the structure of the input text and brought up into a suitable form for further processing.

Database creation

For the current application, it was created a database of articles published between 2000 and 2019 in the Pharmaceutical Research Journal, the official publication of the American Association of Pharmaceutical Scientists. It comprises information like title, abstract, keywords, and funding acknowledgements for a number of 2786 articles.

Pre-processing activities

This step is usually taking the most time of any data-driven project and the current study is not an exception comprising the following steps: 1) documents cleaning 2) text standardization 3) Document-Term Matrix creation 4) TF-IDF Matrix transformation.

Documents cleaning

In the cleaning step, the text is cleaned by removing punctuation and turning it into lowercase letters. This is necessary in order to be able to extract unique terms in a later step, because

[2] https://docs.python.org/3/library/urllib.html
[3] https://pypi.org/project/pyquery/

the computer does not consider equivalent the words "peptide" and "Peptide". In this step the commonly used words like “on, in, of, at etc.” are deleted as they are not carrying any informational value for our task. Additional words that are common for the entire domain are as well deleted, like “pharmaceutical, study, aim, purpose, contribution” etc.

Text standardization

In order to obtain uniqueness of terms, words must be brought to their original form (e.g. nominative for nouns and infinitive for verbs). This can be done by two methods: stemming and lemmatization. Stemming eliminates the suffixes that appear after the word inflection and bring words to the same stem (root) even if the root is not a valid word itself. On the other hand, the approach lemmatization is taking is to reduce the inflected words in a proper way by bringing them to their canonical or dictionary form (Balakrishnan and Lloyd-Yemoh, 2014). For these tasks are used programming packages with embedded language models.

Document-Term Matrix creation

After obtaining the unique roots, a common dictionary is created for all the documents. In our case, it was obtained a dictionary of 21 497 relevant terms. Using this dictionary, we compute the document-term matrix. The obtained object is a mathematical *MxN* sized (M documents and N words in the dictionary) matrix that describes the frequency of terms in a collection of documents. Values of the matrix are just absolute, relative or logarithmic frequencies of the term within each document, denoted with *tf* and calculated with:

$$tf_{w_n,d_m} = \begin{cases} \log\left(1 + f_{w_n,d_m}\right), \text{if } f_{w_n,d_m} > 0 \\ 0, \text{ else} \end{cases}, \quad (4)$$

where f_{w_n,d_m} is the number of occurrences of the term w_n in document d_m, so that we have one *tf* value for each document-term pair.

In short, the *tf* frequency represents how popular a word is in a given document. In general, the higher the frequency, the higher probability the word is more representative for that specific document. The exception are the general terms that are used everywhere, in all documents. Some examples are language specific words (e.g. “and”, “or” etc). These would always get a high rank (high *tf* values) in all the documents without providing much useful insights about content.

TF-IDF Matrix transformation

As a result, by itself, the *tf* metric does not capture the most relevant words for a specific document. In order to downrank the common terms a second concept is needed: the inverse document frequency *idf* (Aizawa, 2003). The inverse frequency measures how popular or unique a term is across documents. It’s computed as the inverse fraction of the documents that contain the word:

$$idf_{w_n,D} = log\,\frac{M}{df_{w_n}} \quad (5)$$

where *D* is the document corpus, *M* is the number of documents in *D* and df_{w_n} is the number of documents containing term w_n. The two components *tf* and *idf* are combined to create the *tf-idf* score:

$$tf.idf\ (w_n, d_m, D) = tf_{w_n,d_m} * idf_{w_n,D} \quad (6)$$

The *tf* value will rank higher the terms occurring often in a document, while *idf* will downrank the common terms. As a result, *tf-idf* will have higher values for terms that are very specific to a given document.

Dynamic Topic Modelling

Validation of number of topics

Now that data is prepared for modelling, *gensim* package for *python* is used to estimate DTM model parameters. The main question here is the number of topics, *K* that should be extracted from data. There are a couple of strategies on how to choose this number: 1) prior domain knowledge (Blei, 2003) 2) topic coherence measure (Newman et al., 2010; Mimno et al., 2011) and 3) manual evaluation by specialists. Because there are no prior domain knowledge nor additional resources like domain experts, the topic coherence measures are used. It is important to keep in mind that under the DTM model conceptualization, all the topics are present in data from any time slice t, but some of them may not be expressed, so that for getting the topic coherence for a couple of alternative K, a static LDA is used on the entire corpus.

In essence, a topic coherence measure is an indicator of how consistent and consolidated topics are inside. Ideally, words are connected to each other through a word chain *(e.g. w1, w2, w1, w4),* so that the word distributions in a topic make sense. In this work, the Mimno et al. (2011) *UMass* intrinsic coherence measure is used:

$$C_{UMass}\left(k, V^{(k)}\right) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} log \frac{df\left(w_n^k, w_l^k\right)+1}{df\left(w_l^k\right)}, \quad (7)$$

where $V^{(k)}$ is the vocabulary given by the static LDA model and is related to *k*th topic, $df\left(w_l^k\right)$ is the number of documents in which w_l is present and $df\left(w_n^k, w_l^k\right)$ is the number of documents that have both terms w_l and w_n. An individual coherence is calculated for each topic and then an average is taken to represent the overall coherence for each of the *K* number of topics used. Based on this approach, the recommended number of topics given by the coherence measures are 85 subjects.

Run DTM with the fixed number of topics

In order to run the DTM algorithm, the same *gensim* package is used. One should give to the function the necessary arguments and objects like the prepared corpus, dictionary, the name of the time dimension variable, number of topics to be considered and the chain variance value. The last argument is used for simulating the dynamic character of both the topic and topic distributions model parameters. It takes a couple of hours to train the DTM model.

At the end of this task, we get a topic assignment for each of the research articles from the database as well and a list that contains the most probable words in each of the topics. Additional to this, there is the possibility of viewing how each topic's terms importance changed over time.

Named Entity Recognition (NER)

Named Entity Recognition (NER) is a key building block of any NLP system. This specific module makes possible the detection and classification of named entities like people, organizations, places in any given text.

The package used for this task it *spaCy*[4], an open-source *python* library used for advanced Natural Language processing. Based on the data presented on the project site, the tool is able to recognize the named entities with a precision of 85%[5].

For extracting the entities, on should pass to the function just the text that contains those names, in our case, articles acknowledgements' sections. The output consists of a list of organization names contained in the given text.

Post-processing activities

The most important task in the part of post-processing is the topic interpretation and description. Of course, not all the topics will make sense for a person that is not having an extensive prior knowledge of the domain and this task is mainly the prerogative of the prepared and experienced researchers. Still, in the interest of system validation, top 10 topics based on the number of published articles were reviewed and interpreted. Identification of topics' name was done through a combination of analysing the most relevant terms and articles' title. The entire dataset of topics' most probable terms and their evolution for each time-slice can be found in the folder of the *github* repository[6].

After extracting the modelling outputs from both DTM and NER modules, results are merged together. This way, for each analysed article we have both assigned interpreted topics and extracted entities.

Results and discussions

Extracted Topics

Based on the DTM module, the extracted corpus from the Pharmaceutical Research Journal was segmented into 85 topics. The allocation of papers to top 5 topics resulted in the distribution illustrated in figure 4. In figure 5 is depicted the yearly evolution of number of articles for each of the top 5 topics.

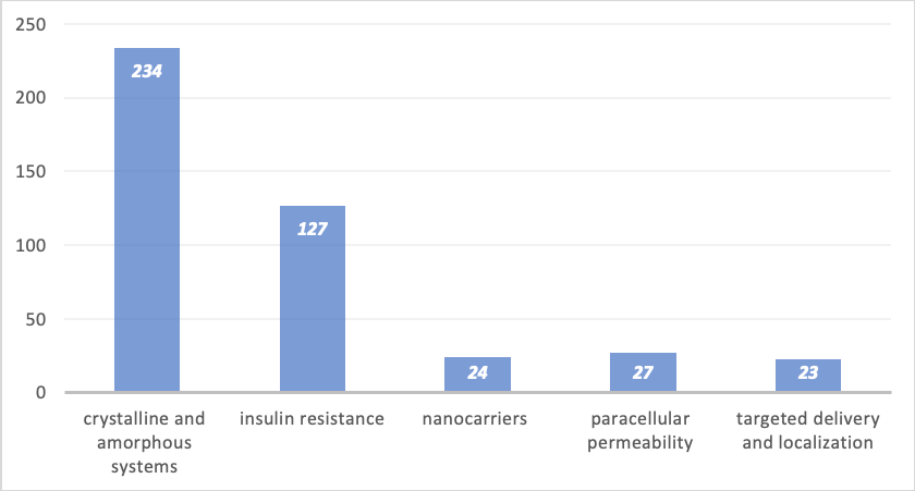

Figure 4. Most popular research areas by number of articles published between 2000-2019

[4] https://spacy.io/

[5] https://spacy.io/models/en

[6] https://github.com/alinkamalvinka/pharma_dtm/blob/master/topics_terms_evolution_all.csv

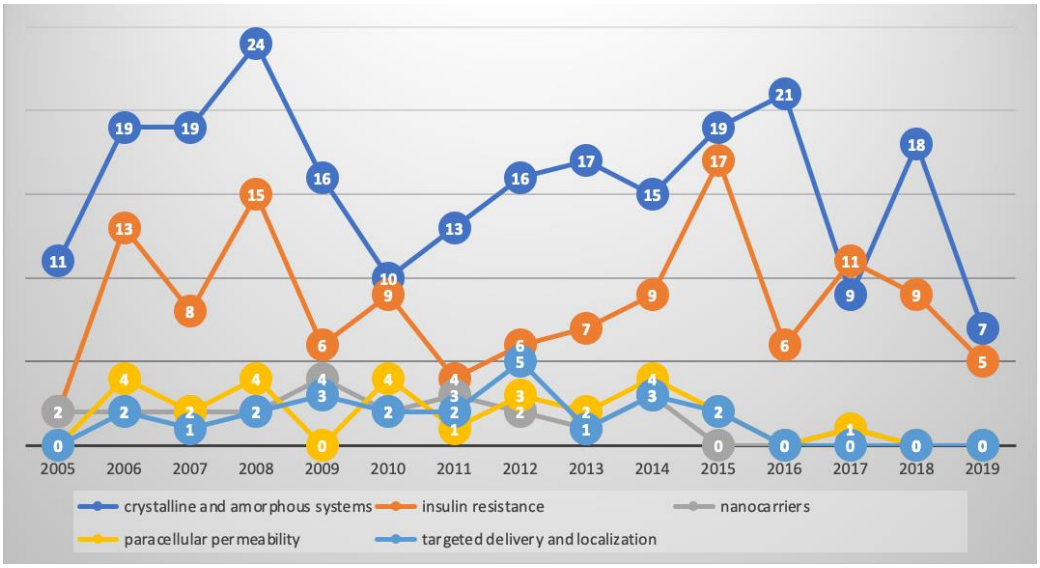

Figure 5. Evolution of the most popular research areas by number of articles published yearly between 2005-2019

For the first research area, crystalline and amorphous systems, we can plot the evolution of terms within the topic itself. Based on the figure 6, we see that at the beginning of the 21st century, the focus was mainly on freeze-drying process to form amorphous solids. We see an increase in importance of doxorubicin as being one of the main drugs for which this technique is used. The period from 2010 until now is characterised through an increased focus on Amorphous formulations for dissolution. This means that the research area is focusing on unlocking new potential and efficacious treatments to combat diseases through solving the poor biopharmaceutical properties, including water insolubility of compounds.

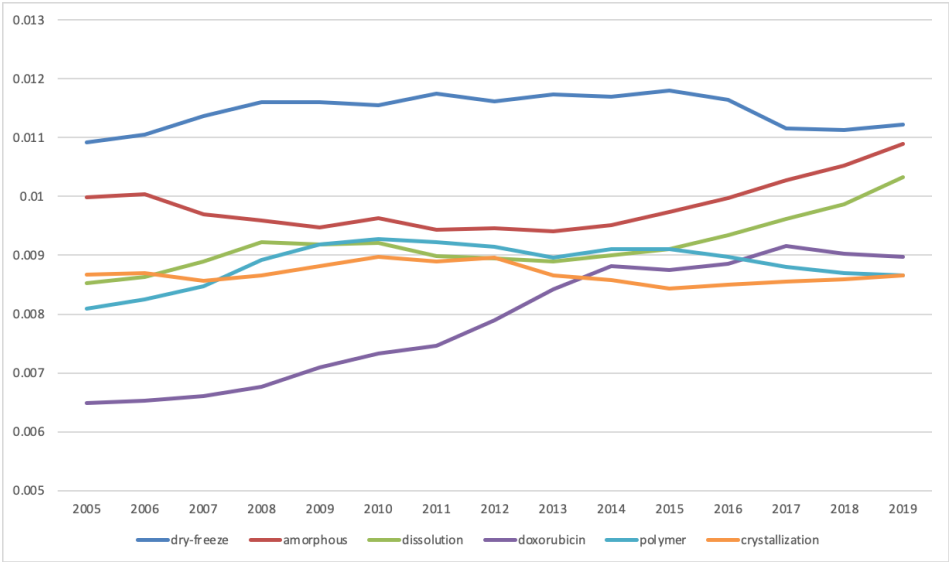

Figure 6. Evolution of the terms for research area of crystalline and amorphous systems

In this research area, efforts are highly segmented. There are a couple of companies that funded researches, like Pfizer, AstraZeneca and Vertex Pharmaceuticals. But as a rule,

the majority of them is done by various national institutes, foundations and university research centres. This emphasise one more time the importance of the secondary research in the drug discovery pipeline.

Conclusion

This paper aimed to design and develop a system that can be used by Pharma scholars and market players in order to analyse the evolution of a research field in a systematic, objective and reliable way. This should enable a more efficient management of the drug discovery process in the early phases. The framework is especially useful to track the development of the area of interest and the competitors' activities w.r.t. the undergoing researches. The reason for this is the functionality of the system that allows new papers to be downloaded, analysed and classified whenever the user needs.

Same time, the framework is able to process a very large number of papers, making the secondary research activities highly time efficient. Another advantage is the transparency of the system. Because the results are reproducible, the analysis can be easily reviewed by other researchers. Additionally, with minimal changes, the system can be used to analyse data from more than one publication.

For the concrete usage of the system on the Pharmaceutical Research Journal articles for the time period from 2000 to 2019, we extracted a number of 85 topics. From these, mostly in focus were crystalline and amorphous systems, insulin resistance, paracellular permeability. When analysing in detail topic evolution for crystalline and amorphous systems research area, it was observed an increased focus on strategies to enhance bioavailability of drugs formulations for dissolution. Also, in this area, the most of research is done by academia and national research centres.

For an improved system, there are two areas that can be addressed. Firstly, when using the system, within organisations there should be a defined process that includes senior and experienced specialists in the loop of deciding relevant journals, number of topics, results interpretation and trends analysis. Secondly, because the DTM cannot detect association between terms if they are not involved in the same document, the obtained results should be linked with other information retrieval tools or proprietary decision systems for a more complete view.

References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45-65.

Alcantara, L. L., Mahichi, F., & Park, Y. (2012). An Analysis of the Antibiotic Industry: An Innovator's Dilemma?. Journal of International Business Research, 11(2), 1.

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data, 6(1), 93.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances.

Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics, 41(5), 706-716.

Bhadury, A., Chen, J., Zhu, J., & Liu, S. (2016, April). Scaling up dynamic topic models. In Proceedings of the 25th International Conference on World Wide Web (pp. 381-390).

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).
Blei, D. M., Franks, K., Jordan, M. I., & Mian, I. S. (2006). Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span. Bmc Bioinformatics, 7(1), 250.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., & Wild, D. J. (2010). Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC bioinformatics, 11(1), 255.
Chen, B., Wang, H., Ding, Y., & Wild, D. (2014). Semantic breakthrough in drug discovery. Synthesis Lectures on the Semantic Web: Theory and Technology, 4(2), 1-142.
Daelemans, W., & Hoste, V. (2002). Evaluation of machine learning methods for natural language processing tasks. In 3rd International conference on Language Resources and Evaluation (LREC 2002). European Language Resources Association (ELRA).
Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Beck, L. (1988, January). Improving information-retrieval with latent semantic indexing. In Proceedings of the ASIS annual meeting (Vol. 25, pp. 36-40). 143 OLD MARLTON PIKE, MEDFORD, NJ 08055-8750: INFORMATION TODAY INC.
Fleming, S. (2018). Pharma's Innovation Crisis, Part 1: Why The Experts Can't Fix It. Forbes Mag.
Frick, J., Guha, R., Peryea, T., & Southall, N. T. (2015). Evaluating disease similarity using latent Dirichlet allocation. BioRxiv, 030593.
Gilbert, J., Henske, P., & Singh, A. (2003). Rebuilding big pharma's business model. IN VIVO-NEW YORK THEN NORWALK-, 21(10), 73-80.
Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. In Advances in neural information processing systems (pp. 17-24).
Groth, P., Loizou, A., Gray, A. J., Goble, C., Harland, L., & Pettifer, S. (2014). API-centric linked data integration: The open PHACTS discovery platform case study. Journal of web semantics, 29, 12-18.
He, B., Tang, J., Ding, Y., Wang, H., Sun, Y., Shin, J. H., ... & Wild, D. J. (2011). Mining relational paths in integrated biomedical data. PLoS One, 6(12), e27506.
Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).
King, G., & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. International Organization, 617-642.
KPMG International Cooperative (2017). Pharma outlook 2030: From evolution to revolution
Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In Advances in neural information processing systems (pp. 121-128).
Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 262-272).
Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. Systematic reviews, 4(1), 172.
Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 100-108).

O'Reilly III, C. A., & Tushman, M. L. (2016). Lead and disrupt: How to solve the innovator's dilemma. Stanford University Press.
Pammolli, F., Magazzini, L., & Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. Nature reviews Drug discovery, 10(6), 428-438.
Philippidis, A. (2015). Despite Big Pharma Retreat, R&D Spending Advances: As Biotechs Fill the Research Gap, Developers of All Sizes Scramble to Reduce Risk. Genetic Engineering & Biotechnology News, 35(06), 6-7.
PricewaterhouseCoopers (PwC) (2009). Pharma 2020: Challenging business models. Which path will you take.
Rizzo, S. J. S., Edgerton, J. R., Hughes, Z. A., & Brandon, N. J. (2013). Future viable models of psychiatry drug discovery in pharma. Journal of biomolecular screening, 18(5), 509-521.
Siebert, M. (2020). How AI and knowledge graphs can make your research easier. Elsevier Connect. See at the URL: https://www.elsevier.com/connect/how-ai-and-knowledge-graphs-can-make-your-research-easier
Stott, K. (2017). Pharma's broken business model: An industry on the brink of terminal decline, Endpoint News, 28 November 2017. See at the URL: https://endpts. com/pharmas-broken-business-model-anindustry-on-the-brink-of-terminal-decline.
Van Vlijmen, H. (2016, March). Open PHACTS: Semantic interoperability for drug discovery. In ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY (Vol. 251). 1155 16TH ST, NW, WASHINGTON, DC 20036 USA: AMER CHEMICAL SOC.
Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., & Wild, D. J. (2011). Finding complex biological relationships in recent PubMed articles using Bio-LDA. PloS one, 6(3), e17243.
Wood, J., Tan, P., Wang, W., & Arnold, C. (2017, April). Source-LDA: Enhancing probabilistic topic models using prior knowledge sources. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE) (pp. 411-422). IEEE.
Xiao, C., Zhang, P., Chaowalitwongse, W. A., Hu, J., & Wang, F. (2017, February). Adverse drug reaction prediction with symbolic latent dirichlet allocation. In Proceedings of the thirty-first AAAI conference on artificial intelligence.