UNIVERSITATEA NAȚIONALĂ DE ȘTIINȚĂ ȘI TEHNOLOGIE POLITEHNICA BUCUREȘTI

FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

DEPARTAMENTUL CALCULATOARE

|  |  |
| --- | --- |
| upb | cs |

PROIECT DE DIPLOMĂ

Titlul proiectului de diplomă (ex: Șablon proiect de diplomă)

Subtitlu (ex: versiunea 2018)

Ioana Popescu

**Coordonator științific:**

Prof. dr. ing. Andrei Ionescu

BUCUREŞTI

2018

UNIVERSITY POLITEHNICA OF BUCHAREST

FACULTY OF AUTOMATIC CONTROL AND COMPUTERS

COMPUTER SCIENCE DEPARTMENT

|  |  |
| --- | --- |
| upb | cs |

DIPLOMA PROJECT

Diploma Project Title (eg: Diploma project template)

Subtitle (eg: 2018 version)

Ioana Popescu

**Thesis advisor:**

Prof. dr. ing. Andrei Ionescu

BUCHAREST

2018

**CUPRINS**

[Sinopsis 2](#_Toc168927500)

[Abstract 2](#_Toc168927501)

[Mulțumiri 3](#_Toc168927502)

[1 Introducere 4](#_Toc168927503)

[1.1 Context 4](#_Toc168927504)

[1.2 Problema 4](#_Toc168927505)

[1.3 Obiective 4](#_Toc168927506)

[1.4 Structura lucrării 4](#_Toc168927507)

[2 Analiza și specificarea cerințelor 5](#_Toc168927508)

[3 STATE OF THE ART 6](#_Toc168927509)

[3.1 YAKE! (Yet Another Keyword Extractor) 6](#_Toc168927510)

[3.2 Latent Dirichlet Allocation 9](#_Toc168927511)

[3.3 spaCy 12](#_Toc168927512)

[3.3.1 The spaCy pipeline 12](#_Toc168927513)

[4 Soluția propusă 13](#_Toc168927514)

[4.1 Indicații formatare formule 14](#_Toc168927515)

[5 Detalii de implementare 15](#_Toc168927516)

[5.1 Indicații formatare tabele 15](#_Toc168927517)

[6 Studiu de caz / Evaluarea rezultatelor 17](#_Toc168927518)

[7 Concluzii 19](#_Toc168927519)

[8 REFERENCES 20](#_Toc168927520)

[9 Anexe 21](#_Toc168927521)

# Sinopsis

Sinopsisul proiectului are rol de introducere, conținând atât o descriere pe scurt a problemei abordate cât și o enumerare sumară a rezultatelor și a concluziilor. Se recomandă ca sinopsisul să fie redactat într-un limbaj accesibil unei persoane nefamiliarizate cu domeniul, dar în același timp destul de specific pentru a oferi rapid o vedere de ansamblu asupra proiectului prezentat.

Sinopsisul proiectului va fi redactat atât în română cât și în engleză. Ca dimensiunea recomandată aceasta secțiune va avea maxim 200 de cuvinte pentru fiecare variantă. Împreună, ambele variante se vor încadra într-o singură pagină.

# Abstract

The abstract has an introductory role and should engulf both a brief description of the issue at hand, as well as an overview of the obtained results and conclusions. The abstract should be formulated such that even somebody that is unfamiliar with the projects’ domain can grasp the objectives of the thesis while, at the same time, retaining a specificity level offering a bird’s eye view of the project.

The projects’ abstract will be elaborated in both Romanian and English. The recommended size for this section is limited to 200 words for each version. Together, both versions will fit in one page.

# Mulțumiri

(opțional) Aici puteți introduce o secțiunea specială de mulțumiri / acknowledgments.

# Introducere

Parametrii de formatare recomandați pentru lucrare:

* Font recomandat: Calibri; Dimensiune font: 12;
* Spațiere între linii: 1,15; Spațiere după paragraf: 8pt;
* Stil: Justified;
* Dimensiune pagină: A4; Margini: 2,54cm/ 2,54cm/ 2,54cm/ 2,54cm;
* Heading1: Calibri, 14, bold, all caps;
* Heading2: Calibri, 14, bold;
* Heading3: Calibri, 12.
* Font pentru formule: Cambria Math, 12.

În cadrul introducerii, este necesară abordarea următoarelor puncte care reprezintă de fapt familiarizarea cititorului (comisia, alți colegi sau experți în domeniu) cu tema proiectului, soluția propusa și cuprinsul/structura lucrării. Deși introducerea poate conține și unele elemente mai generale, se recomandă păstrarea unui limbaj tehnic, specific audienței care va citi lucrarea.

În cadrul capitolelor următoare, veți regăsi o serie notații de forma [Dezvoltare de produs], [Cercetare]. Acest tip de formatare este utilizat exclusiv în acest template pentru a marca sfaturi și cerințe specifice pentru lucrări de diploma cu specific diferit. În pregătirea documentului vostru, nu veți utiliza aceste marcaje.

Elementele pe care trebuie să le abordați în introducere sunt descrise în cadrul subcapitolelor de mai jos.

## Context

O scurtă introducere a proiectului, motivație, explicație de ce este relevant domeniul proiectului.

## Problema

Care este problema pe care proiectul o va rezolva.

## Obiective

Care sunt obiectivele proiectului/soluției/abordării/ideii; Ce creșteri sau evoluții determină rezolvarea proiectului.

## Structura lucrării

Un paragraf în care fiecare dintre secțiunile următoare este prezentată în 1-2 fraze, punând accentul pe elementele cele mai semnificative din fiecare secțiune.

# Analiza și specificarea cerințelor

[Dezvoltare de produs] Acest capitol va analiza cerințele produsului din prisma potențialilor clienți și a scenariilor de utilizare preconizate, urmând a fi generată o lista de funcționalități.

[Cercetare] Acest capitol va introduce motivația realizării proiectului propus.

Dacă proiectul de licență face parte dintr-un proiect mai amplu (de exemplu un proiect complex, la care lucrează 2 studenți (ex: 1 student la front-end-ul aplicației, 1 student la back-end-ul aplicației), în acest capitol va fi explicat pe scurt ansamblul proiectului și ce parte din proiect este adresată de lucrarea propusă.

Criterii pentru calificativul Nesatisfăcător:

* [Dezvoltare de produs] Cerințele sunt imaginate de student pe baza unei analize a pieței;
* [Cercetare] Nu se oferă o motivație valida.

Criterii pentru calificativul *Satisfăcător*:

* [Dezvoltare de produs] Există un interviu, un client, analiza cerințelor este elaborată pe baza interviului;
* [Cercetare] Motivația este doar personala.

Criterii pentru calificativul *Bine*:

* [Dezvoltare de produs] Proces iterativ pe baza unor interviuri cu mai mulți clienți, dezvoltare MVP, reevaluare cerințe;
* [Cercetare] Motivația este legata de o necesitate științifica / tehnica explicită.

# STATE OF THE ART

In the field of Natural Language Processing (NLP), keyword extraction is essential for text mining purposes, such as automatic indexing, automatic summarization, automatic classification, automatic clustering, automatic filtering and topic detection (Zhang, 2008). The necessity of accurately extracting keywords and topics from documents has led to the development of numerous algorithms. However, this area is still not as precise as other computer science fields (Campos et al., 2018). There is no universal approach for effectively extracting semantic information without requiring either large manually labelled data sets for training or a comprehensive preprocessing of the text, according to the specifics of the domain and of the individual document. Therefore, one needs to use more than one tool in order to achieve the expected results. In this chapter, I will discuss the tools integrated in my project: YAKE! (Yet Another Keyword Extractor) for keyword extraction, LDA (Latent Dirichlet Allocation) for topic modelling, the spaCy Python library for preprocessing, part of speech tagging and named entity recognition, fastText for language detection and the Gensim package for the implementation of LDA.

## YAKE! (Yet Another Keyword Extractor)

Supervised methods have been the prevailing approach in keyword extraction, but they require extensive training and manual work for annotating large collections of documents (Campos et al., 2020). Moreover, these machine learning models may not perform well on domains that they were not trained on.

Despite being an unsupervised method, TF-IDF (Term Frequency-Inverse Document Frequency) also needs an extended collection of documents since it assumes that a term is relevant in a specific text if it occurs frequently in that document but is otherwise rare across the collection. The formula used to measure the significance of a word is (Nomoto, 2022):

(1)

The first term in equation (1) is the Term Frequency (frequency of word in document j), while the second term is known as the Inverse Document Frequency, where the numerator represents the total count of documents and the denominator is the number of documents that contain word (Nomoto, 2022).

RAKE (Rapid Automatic Keyword Extraction) is another unsupervised keyword extractor that makes use of a co-occurrence matrix, thus rewarding with a higher score the words that appear not only frequently, but also in longer phrases (Rose et al., 2010):

(2)

In formula (2), the degree (deg) is the total length of the potential keyphrases that word w appears in, and the frequency represents the count of w in those sentences.

The score of a candidate keyphrase is computed by adding up the scores of the component words.

The subsequent unsupervised algorithm is YAKE, the approach that I chose for my implementation. YAKE ranks candidate keywords by calculating a score based on statistical text features. It does not rely on a pre-labelled corpus and it is equally effective regardless of domain or language, only requiring a stop words list specific to the language it is applied on (Campos et al., 2020). YAKE’s ability of operating on individual documents without the necessity of additional corpora makes it suitable for extracting keywords from a set of technical abstracts, a continuously expanding and changing collection (Rose et al., 2010).

O imagine care conține text, captură de ecran, Font, document

Descriere generată automat

Figure 1: YAKE Algorithm (Campos et al., 2020)

Figure 1 presents YAKE as pseudocode, highlighting the five steps of the algorithm (Campos et al., 2020):

1. Text pre-processing and candidate term identification

The text is split into sentences, sentences are split into chunks delimited by punctuation marks and chunks are divided into tokens. Then, stop words are removed and tokens are tagged with the following labels: d (digit or number), u (unparsable), a (acronyms), u (uppercase), p (parsable).

1. Feature extraction

The algorithm favours tokens that appear repeatedly in uppercase or abbreviated as an acronym and assumes a term is more important if it can be found closer to the beginning of the text. Therefore, it proceeds by calculating some statistics for each term: the frequency, the number of times that the word occurs capitalized inside sentences, the number of times that the term appears as an acronym and the positions of the sentences that contain it in the document.

Afterwards, YAKE calculates five features for each token and combines these attributes into the following formula which quantifies the significance of a token:

(3)

In formula (3), evaluates the casing dimension, taking into account the frequency of occurrence with an initial capital letter and the frequency of the acronym. The factor measures the importance of a word based on all of its positions in the text, more precisely the indices of the sentences it occurs in. The “term frequency normalization” () represents the frequency of a word, normalized to ensure that, in large documents, this element does not overweigh the other features in the formula. The fourth component, “term relatedness to context” (), aims to downgrade the words that have similar characteristics to those of stop words, specifically a high number of occurrences with many different words around them. The last feature in the formula is (“term different sentence”), which rewards tokens that appear in numerous sentences.

1. Computing term score

The score of each token is determined using formula (3), the most significant terms having the lowest score.

1. Generating n-grams and calculating the scores of potential keyphrases

In the fourth stage, the YAKE algorithm finds potential keywords by extracting n-grams (with the maximum n given as input to the algorithm) from the chunks obtained in the first step, choosing only the phrases that do not contain tokens marked as “digit” or “unparsable” and do not start or end with a stop word.

The score of a keyword kw is determined as:

(4)

In equation (4), the multiplied scores of the component tokens of kw are divided by the sum of these scores amplified by the frequency KF of the keyword. Consequently, the algorithm can differentiate between potential keywords that contain the same words reordered and choose the most frequent one.

1. Handling duplicates

After ranking the keywords by the lowest score, YAKE removes duplicates using a similarity measure, which can be either the Levenshtein similarity, the Jaro-Winkler similarity or the sequence matcher. Two keywords are considered duplicates if their distance, computed using one of the three previous methods, surpasses a deduplication threshold chosen before starting the algorithm and given as parameter. When two keyphrases exceed this similarity threshold, the one with the higher score is removed from the final keyword list.

## Latent Dirichlet Allocation

Topic modeling methods aim to discover “hidden” (latent) topics from large collections of documents. Each topic consists of a set of words that frequently occur together and thus it is assumed that they are conceptually related. Given its capacity of capturing semantic features of corpora, topic modeling is an effective tool in text mining and information retrieval, used in text summarization and sentiment analysis applications (Abdelrazek et al., 2023).

There are four categories of topic models: algebraic, fuzzy, probabilistic and neural (Abdelrazek et al., 2023).

The most notable algebraic topic modeling method is Latent Semantic Indexing (LSI) (Deerwester et al., 1990), that represent a collection of documents as a term-document matrix, where an element of the matrix is the frequency of a word in a document (a row corresponds to a word, while documents are placed on columns). Deerwester et al. (1990) decomposed the term-document matrix using singular value decomposition (SVD) in order to obtain a vectorial representation for words and texts. Geometrically, the distance (calculated as cosine similarity) between these vectors indicates the degree of semantic correlation between two terms, two documents or a term and a document. Although LSI is proficient in capturing synonymy, it does not handle polysemy with the same efficacy (Deerwester et al., 1990).

LSI is not based on a solid statistical foundation and a flaw that it presents is the false assumption that words and documents follow a joint Gaussian distribution (Abdelrazek et al., 2023). These concerns are solved by the probabilistic approaches, which, according to Abdelrazek et al. (2023), dominated the field before neural topic models began their rising around 2015.

Latent Dirichlet Allocation (LDA), a generative probabilistic model developed by Blei et al. (2003), remains one of the most popular topic modeling methods. LDA is characterized by a series of assumptions:

* documents and words are exchangeable, their order does not matter, concept known as the bag of words model (Blei et al., 2003);
* each document is a probability distribution over topics and the distributions of topics in all documents share the same Dirichlet prior (Jelodar et al., 2017);
* each topic is a distribution over words and the distributions of words across topics are characterized by a common Dirichlet prior (Jelodar et al., 2017).

The generative process of LDA begins by choosing the Dirichlet priors α and β, corpus-level parameters. Afterwards, for each document, the algorithm chooses the distribution θ of topics in the current document from a Dirichlet distribution of parameter α. Then, for each of the N words of the document, a topic is sampled from distribution θ and a word is chosen from the distribution of this topic (Blei et al., 2003). Thus, the generative process of LDA iteratively calculates the probability distribution of topics in documents and the distribution of words within topics, by randomly assigning them based on a probability distribution, until the results converge.

Griffiths & Steyvers (2004) used LDA to extract topics from scientific abstracts and achieved meaningful results, as in the following figure:

O imagine care conține text, Font, alb, chitanță

Descriere generată automat

Figure 2: An abstract with the words labelled according to the topic they belong in

In figure 2, the words in an abstract are tagged with a number corresponding to the topic they belong in. The highlighted words are extracted from the topic with the highest probability, hence they can be considered as keywords of the abstract. According to Griffiths & Steyvers (2004), these words accurately summarize the content of the document.

The limitations of LDA stem from the variability of its parameters which need to be selected before running the algorithm: the Dirichlet priors α and β and the number of topics. Although these parameters can be fine-tuned using variational inference or Gibbs sampling, their variations significantly impact the performance of the algorithm (Griffiths & Steyvers, 2004). The value of α determines the number of topics being assigned to a document (a higher value leads to more topics), while the value of β controls the sparsity of the topic-word distributions (a higher β determines the algorithm to assign similar probabilities to more words). Regarding the number of topics, a small number may lead to underfitting, while a large number of topics generally results in overfitting the data (Tijare & Rani, 2020).

Another considerable issue is that LDA requires a thorough clean-up and preprocessing of the documents; stop word removal and lemmatization are very important, otherwise stop words will occur among the words that define topics and different forms of the same word will be considered separately (Moreno-Ortiz, 2024).

Tijare & Rani (2020) compared the results of LDA on social media data, using the implementations from popular Python packages Gensim and scikit-learn, and visualizing them with pyLDAvis, a tool which graphically represents topics. They also computed a coherence score for the topics, a measure of their relevance.

O imagine care conține linie, Interval, diagramă, captură de ecran

Descriere generată automat

Figure 3: Coherence score for scikit-learn LDA (Tijare & Rani, 2020)

Figure 3 shows the coherence score depending on the number of topics. Despite the score rising, Tijare & Rani (2020) observed that topics begin to overlap when there are more than 10 of them.

O imagine care conține captură de ecran, cerc, diagramă, proiectare

Descriere generată automat

Figure 4: pyLDAvis representation of LDA results obtained with Gensim

In figure 4, which depicts 10 topics extracted using the Gensim implementation of LDA, topics 7, 9 and 10 slightly intersect, indicating they share common terms.

Tijare & Rani (2020) also compared LDA with LSI and NMF, another topic modeling technique, and concluded that LDA achieves better coherence scores. They used two methods for computing a coherence score, with U\_MASS resulting in a coherence of for LDA and for LSI (negative values suggest coherence; the lower the value, the more coherent the topics) and C\_V showing a score of for LDA and for LSI (LDA has the higher score, thus the higher coherence).

## spaCy

spaCy is a state-of-the-art natural language processing Python library, that offers a customizable processing pipeline, with components capable of tokenization, lemmatization, recognizing named entities, tagging parts of speech and dependency parsing (*SpaCy*, n.d.):

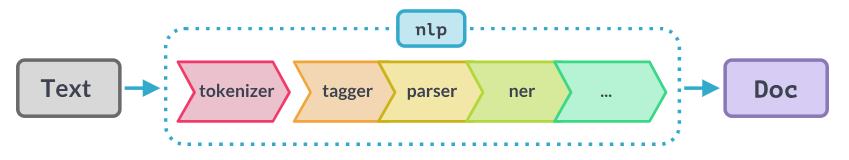


Figure 5: spaCy pipeline

Figure 5 shows the default components of the spaCy pipeline, which can be modified by disabling some components or by adding custom components, developed by the user or available in the open-source community.

The motivation behind spaCy is explained by its creator, Matthew Honnibal, who wanted to develop a fast, efficient and compact language processing tool, suitable for production environments, as opposed to the older libraries, which he believes offer too many features that are otherwise nor effective nor up to date with the advancements in computer science (Matthew Honnibal, n.d.). Honnibal explained why he did not contribute to popular NLP library NLTK instead of starting a new project, stating that he thinks the maintainers of NLTK should not keep adding to the project, but “throw almost all of it away”.

The reason why numerous NLP packages, NLTK included, became so large yet not industry-grade is simply because this is not their target, given the fact that they are developed by researchers for researchers and students (Srinivasa-Desikan, 2018).

In the following sections, I will detail the default components of the spaCy pipeline, specifically the components of the trained model that I used in my project, “en\_core\_web\_lg”, a large general-purpose pipeline, trained on web articles, with 514,000 word vectors (*SpaCy*, n.d.).

O imagine care conține text, captură de ecran, diagramă, linie

Descriere generată automat

Figure 6: en\_core\_web\_lg pipeline (*SpaCy*, n.d.)

The components of the common spaCy models are pictured in figure 6, along with their dependencies.

### The Tokenizer

The tokenizer is not present in the figure above, since it is considered different from the other components in that it receives a string as an input and returns a Doc object, while the others receive the Doc and return a processed version of it. Despite this, the tokenizer is an essential pipeline step, and it is the first to run, before the tok2vec component in the case of the en\_core\_web\_lg model.

The Doc returned by this pipeline element contains the tokens obtained by splitting the text into words, whitespaces and punctuation marks. The tokenizer treats special cases according to the language, for instance it splits “Let’s” or “don’t” into two tokens even though they are not delimited by a whitespace, while it considers acronyms like “U.K.” and “N.Y.” as single tokens.

O imagine care conține text, captură de ecran, Font, număr

Descriere generată automat

Figure 7: spaCy tokenizer

Figure 7, taken from spaCy’s official documentation (*SpaCy*, n.d.), shows how spaCy tokenizes a sentence: it splits it by whitespaces first, then it treats special cases such as exceptions (“Let’s” and “N.Y.”), prefixes (opening quotes) and suffixes (exclamation point, closing quotes) from left to right.

### The tok2vec

The token-to-vector component transforms the tokens into word vectors, numerical representations that encapsulate the semantics of words. Other components that rely on word embeddings can use the results made available by the tok2vec layer through listeners, as it can be seen in figure 6 in case of the tagger, the parser and the lemmatizer. However, the named entity recognizer (ner) is intended to be independent of the other steps so that it can be used by itself, hence it contains its own tok2vec layer.

### The part of speech tagger

The POS tagger makes statistical predictions about the syntactic role of each token. The Token class contains two attributes that provide information about the part of speech: “pos\_” and “tag\_”, which goes into more detail. I will illustrate the tagger using the same sentence from figure 7.

Table 1: POS tags

|  |  |  |
| --- | --- | --- |
| TOKEN | POS | TAG |
| “ | PUNCT | punct |
| Let | VERB | VB (verb, base form) |
| ‘s | PRON | PRP (pronoun, personal) |
| go | VERB | VB |
| to | ADP (adposition) | IN (conjunction, subordinating or preposition) |
| N.Y. | PROPN (proper noun) | NNP (noun, proper singular) |
| ! | PUNCT | punct |
| “ | PUNCT | punct |

I obtained the results in table 1 by applying the spaCy pipeline on the chosen sentence and printing the pos\_ and tag\_ attributes of each token, as well as the explanation of what the abbreviations mean, using spacy.explain().

### The dependency parser

The dependency parser builds a dependency graph that models the syntactic structure of a phrase, where words are represented as nodes and the edges are labelled with the relationships between words.

O imagine care conține schiță, diagramă

Descriere generată automat

Figure 8: displaCy dependency visualizer

The following dependencies are pictured in figure 8, which I made using displaCy visualizer[[1]](#footnote-1):

* ccomp - clausal complement
* nsubj - nominal subject
* prep - prepositional modifier
* pobj - object of preposition

### The attribute ruler

The attribute ruler layer allows users to specify some rules for setting token attributes that may be incorrectly handled by other pipeline components such as POS tagger, ner or lemmatizer. This step is suitable in case of domain specific contexts or unusual uses of some words.

### The lemmatizer

The lemmatizer assigns base forms to tokens based on the POS tag, morphological rules and dictionary lookups. For instance, the lemma of the “’s” token in the previously analysed sentence “Let’s go to N.Y.!” is “us”, while the other lemmas coincide with the initial words.

### The named entity recognizer

The entity recognizer is a neural network model capable of annotating different types of proper nouns and numerical values with the following labels: CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK\_OF\_ART (*SpaCy*, n.d.).

O imagine care conține text, captură de ecran, Font, număr

Descriere generată automat

Figure 9: displaCy entity visualizer [[2]](#footnote-2)

Figure 9 shows an example of a text labelled by the spaCy named entity recognizer.

According to the spaCy documentation, the part of speech tagger in en\_core\_web\_lg has an accuracy of 0.97 and its named entity recognizer has a precision of 0.85, a recall of 0.86 and an F-score equal to 0.85.

## fastText

fastText is a linear text classifier developed by Facebook AI Research (Joulin et al., 2016), that achieves an accuracy similar to that of deep neural networks classifiers, while being remarkably faster. The training of fastText lasts less than ten minutes for more than one billion words and the classification of 500,000 sentences takes less than one minute on an average multicore CPU.

fastText represents the text as a bag of n-grams in order to retain some information about the order of words, that the simple bag of words model does not consider. Words and n-grams are converted into word vectors.

During training, the probability distribution over classes is computed using hierarchical softmax. The classes are organized in a binary decision tree and the probability of a class is computed by multiplying the probabilities of all the nodes visited during a depth first traversal from root to the leaf node which corresponds to that class. The purpose of training is to learn the configurations of matrixes A and B that minimize the loss between the predicted classes and the actual labels (Joulin et al., 2016):

(5)

In formula (5), N is the number of documents, is a vector which indicates the correct label of document n ( if document n belongs in class k, otherwise ), is the normalized vector of n-gram features and f is the softmax function described above, thus is the probability distribution over classes for document n. The algorithm uses stochastic gradient descent to find the parameters A and B that minimize the value of the negative log-likelihood formula (5) (Joulin et al., 2016).

Joulin, Grave, Bojanowski, Douze, et al. (2016) described some model compression techniques that can drastically reduce the memory usage without greatly impacting the accuracy: product quantization (splitting the word vectors into smaller subvectors and then quantizing them using nearest neighbour), pruning the vocabulary (keeping only a fixed number of the most relevant words and n-grams), hashing trick and Bloom filter (hashing words and n-grams in order to reduce the size of the dictionary and using Bloom filters to check the presence of a term in the dictionary in constant time).

I used the Python fastText library for the language identification purpose[[3]](#footnote-3). It provides two models capable of recognizing 176 languages, one with a size of 126MB and the compressed version with a size of 917kB. I chose to use the larger model, because it offers slightly superior speed and accuracy.

## Gensim

Gensim is an open-source natural language processing Python library[[4]](#footnote-4) focused on topic modeling, specifically unsupervised semantic modeling methods like the ones previously discussed in this paper: Latent Semantic Indexing/Analysis and Latent Dirichlet Allocation. Radim Řehůřek, the creator of Gensim, acknowledged the lack of easy to use and efficient NLP libraries[[5]](#footnote-5) and committed to bridging the “gap between academia and ready-to-use software packages” (Řeh uřek & Sojka, 2010).

Gensim is based on two fundamental ideas, according to Řeh uřek & Sojka (2010): document streaming, which guarantees memory independence by not storing the whole large corpus in RAM, and transformations between vector spaces. Gensim relies on the Vector Space Model, the paradigm of representing documents as vectors of features. This model allows for easy estimation of document resemblance through cosine similarity and facilitates the transformation of documents from one model representation to another (for instance from TF-IDF to LSA) by performing a vector space translation.

I used the LdaMulticore model from Gensim, a parallelized version of the LDA algorithm, capable of significantly reducing the training time[[6]](#footnote-6):

Table 2: LdaMulticore and LdaModel training time

|  |  |
| --- | --- |
| Algorithm | Training time |
| LdaMulticore(workers=1) | 2h30m |
| LdaMulticore(workers=2) | 1h24m |
| LdaMulticore(workers=3) | 1h6m |
| old LdaModel() | 3h44m |
| iterating over the corpus | 20m |

Table 2, taken from the Gensim documentation, shows the duration of training LdaMulticore and the old LdaModel on all the English articles from Wikipedia, measured on an i7 server with 4 physical cores. After deducting the 20 minutes spent iterating over the documents, it can be concluded that LdaMulticore, even when it runs on a single thread, is almost 3 times faster than the basic LdaModel, while with 2 workers it achieves a speed-up equal to 6 and with 3 cores the speed-up reaches 8.35.

# proposed solution

The proposed solution integrates a variety of natural language processing techniques and tools that include statistical, probabilistic and machine learning methods, both supervised and unsupervised. Together, along with a carefully thought-out preprocessing, tailored to the specific NLP tools employed and the collection of data to be processed, they form an effective system, capable of discovering themes from researchers’ abstracts. I implemented two different methods to achieve this: keyword extraction with YAKE and topic modeling using LDA, two approaches that complement each other and manage to extract meaningful words that reveal the sphere of specialization for any author.

I chose to develop my solution using the Python programming language, the most suitable for natural language processing applications for its speed, versatility and collection of the most powerful NLP libraries, as described in chapter 3.

In the current chapter, I will outline the architecture of my solution, which includes three main steps for obtaining the keywords from a researcher’s abstracts: cleaning up the abstract collection, extracting keywords with YAKE, which contains another preprocessing step at the abstract level, and extracting significant words with LDA, which also integrates a clean up stage.

## Filtering the abstract list

Before preprocessing the abstracts and extracting themes from them, it is important to remove the texts that do not contain any useful information, to prevent them from negatively impacting the results and to reduce the unnecessary use of memory and computational resources.

During this step, the program takes an author’s list of abstracts, extracted from the research platform, and first filters out the empty abstracts or the abstracts that are not composed of words. Among the top 20 authors on the platform, with the most publications, there were 3416 abstracts extracted from the database that were either null or did not contain any alphabetic characters.

From the non-empty texts that contain letters from the English alphabet, abstracts that are not written in English are removed, using the fastText library for language prediction. Besides the fact that the presence of these non-English abstracts in the corpus would unnecessarily take up resources, it would also impact the results. The English spaCy pipeline cannot parse and label non-English texts correctly and the YAKE algorithm is not able to handle this kind of documents either, since it relies on a static stop word list to identify the irrelevant terms. With fastText, I found 269 non-English abstracts among the top 20 researchers. Given the fact that LDA computes a higher probability of appearance within a topic for the more frequent words in the corpus, this means that there is a high chance that some of the most frequent non-English words will appear among the keywords extracted with LDA. For instance, I tried applying LDA to the abstracts of the author who has the most abstracts not written in English, to extract one topic, without priorly removing the non-English texts, and the Romanian preposition “de” was among the top 15 words.

The next step in the corpus preprocessing entails the use of spaCy for the named entity recognition (NER) pipeline, in order to remove the abstracts that contain more person and organization names than useful information, step that I introduced after discovering that 77 of the input texts were nothing more than an enumeration of personal names and institutions. The spaCy named entity recognizer does not always label those names correctly, especially given the fact that they are not part of a clear context, for instance:

O imagine care conține text, captură de ecran, Font, număr

Descriere generată automat

Figure 10: Named entity recognizer on input text

Figure 10 represents the output of the displaCy visualizer and shows an example of a text in which the spaCy NER does not label all the named entities properly. “Mihnea COSTOIU” is one of the names incorrectly not labelled as “PERSON”, and it appears among the keywords extracted with YAKE:

O imagine care conține text, captură de ecran, Font

Descriere generată automat

Figure 11: YAKE output with person names

Figure 11 shows two researchers’ names in the output of YAKE, as proper nouns are more likely to be extracted by YAKE, because it considers capitalized words more important.

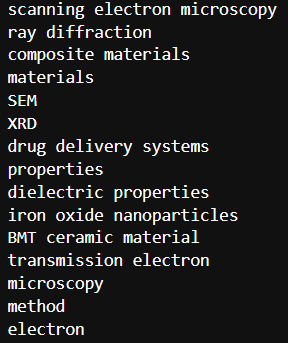


Figure 12: YAKE output without names

In figure 12, the output of YAKE for the same author, after removing those input texts, the two personal names do not occur among the keywords anymore. Therefore, it is essential to filter out texts that contain mostly names before extracting keywords.

## Extracting keywords with YAKE

The next step of the proposed solution is the keyword extraction using YAKE. This algorithm does not require a comprehensive preprocessing, since it uses punctuation and stop words to identify candidate keyphrases. However, I removed some words that I did not want to appear in the resulting keyword list.

### Preprocessing

Firstly, before extracting the keywords, I concatenated all the abstracts of a researcher into a single string, because YAKE operates on individual documents. For preprocessing the resulting text, I used the spaCy NER pipeline component to remove the following named entities: personal names, nationalities, religious or political groups, buildings, geopolitical and other types of locations, date and time entities, monetary values and other numerical figures. I did not eliminate organizations, as I observed that a lot of relevant terms that were written with initial uppercase letters were frequently classified by spaCy as organization names, such as “Big Data”, “Cloud Computing” or “Natural Language Processing”.

O imagine care conține text, captură de ecran, Font, proiectare

Descriere generată automat

Figure 13: Significant terms incorrectly classified as ORG

Figure 13 pictures an abstract of one of the top 20 authors, where the “Big Data” term is labelled seven times as an organization, suggesting how significant concepts like this would be lost if I chose to remove “ORG” entities during preprocessing. To illustrate this, I extracted keywords from the publications of the same author, adding organizations to the list of entities to remove, and I obtained the keyphrases below:

O imagine care conține text, captură de ecran, Font

Descriere generată automat

Figure 14: keyword results after removing ORGs

In figure 14, “big data” appears once as a keyword and once inside another keyphrase, but it only appears in lowercase and at the bottom of the keyword list.

O imagine care conține text, captură de ecran, Font

Descriere generată automat

Figure 15: keyword results without removing ORGs

On the other hand, when I kept the entities labelled as organizations in the text, “Big Data” appeared in three keyphrases, two of them being in the top 3 most relevant keywords, as we can see in figure 15.

Nevertheless, there are some institution names that should be removed, because they are very frequent in these academia writings, such as names of universities and faculties. I prevented their occurrence among keywords by adding the words “university” and “faculty” to the YAKE’s internal stop word list, along with numerous other words specific to technical and scientific papers, that I encountered many times in keyword extraction results, such as “conference”, “workshop”, “experiment”, “professor” and “research”.

### Keyword extraction

After preprocessing the text, the program proceeds with the keyword extraction using the yake Python library[[7]](#footnote-7). I extracted 15 keywords for an author, because I think this number is enough to summarize the research activity of an author, without being so large that meaningless keyphrases are extracted, especially in the case of researchers with fewer publications.

O imagine care conține text, captură de ecran, Font

Descriere generată automat

Figure 16: Result of YAKE with 20 keywords

Figure 16 shows the results of my attempt to extract 20 keywords from the abstracts of an author who has only 10 publications. Starting from the fourteenth keyword, there are some phrases that are not only irrelevant as keywords, but also meaningless from a grammatical perspective, for example “imposed major penalties” and “process planning consists”.

Campos et al. (2020), the creators of YAKE, observed that the vast majority of keywords extracted by humans (gold keywords or ground-truth keywords) are unigrams, bigrams and trigrams. They experimentally obtained the best results with YAKE when extracting n-grams with the maximum value for n being equal to 3. Thus, I also chose trigrams as the maximum n-gram size in my solution.

For the similarity measure, I used the default option of the YAKE implementation, the Levenshtein distance. According to Campos et al. (2020), when using the Levenshtein similarity, the algorithm gives the best results if the widow size is equal to 1 (the number of words before and after a term which are counted when determining the number of words this term co-occurs with) and the deduplication threshold has a value of 0.8. However, for the purpose of my project I thought a threshold equal to 0.5 leads to the best results, because a bigger value produces too similar keywords and other relevant terms are lost, while a smaller value causes some significant keyphrases to be excluded because of their similarity to other keywords. A lower deduplication threshold may be particularly problematic in the case of researchers with a few abstracts. Instead of selecting relevant keyword candidates with high similarity, the algorithm may extract some meaningless keyphrases in its attempt to remove duplicates.

O imagine care conține text, captură de ecran, Font, software

Descriere generată automat

Figure 17: Comparison between deduplication threshold values

Figure 17 compares the keyphrases extracted for a top 20 author with 380 papers, with three different values for the deduplication threshold. On the left, there are the keywords obtained for a low threshold of 0.2, which include some frequent but not very significant words like “data”, “system”, “information” and “development”. The terms on the right side, which resulted after setting the threshold to 0.8, are too similar, with the word “language” appearing in 9 out of 15 keywords, in approximately the same context (“Natural Language Processing”, “Natural Language”, “Language Processing” etc.). Therefore, a high deduplication threshold produces some redundant keywords and excludes some relevant ones (for instance “neural network model” and “Latent Semantic Analysis”), that are captured by choosing a 0.5 threshold, as pictured in the centre of the figure above.

After setting these parameters, the program calls the keyword extraction function from the yake package and the 15 keyphrases are returned and printed.

## Topic modeling with Latent Dirichlet Allocation

The second tool that I used for extracting meaningful themes from research abstracts is Latent Dirichlet Allocation. Despite keyword extraction not being the purpose of LDA, which is typically used for indexing, summarization or classification, it can be successfully employed as a keyword extractor. The most representative words of the uncovered topics can be treated as keywords of the document collection, as long as the preprocessing step brings the corpus to an adequate form.

### Preprocessing

Preprocessing plays a critical role when working with LDA, therefore this time I used the whole spaCy pipeline, not only the named entity recognizer. Firstly, stop word removal is especially important. Given the generally high frequency of stop words, their probabilities of occurring within topics may surpass that of meaningful terms. I removed the stop words from the default spaCy list, as well as the custom stop words that I discovered, the same used for YAKE. The named entity removal is also similar to the implementation described in the previous chapter about YAKE, as I removed the same entities.

Justeson & Katz (1995) studied the particularities of technical and scientific terminology and found that approximately 95.6% of terms are noun phrases. Among 800 terms, they found only 32 adjectives and 3 verbs, the rest of them being noun phrases. Thus, I used the part of speech tags assigned by the spaCy POS tagger to keep only the nouns and the adjectives, removing all the other parts of speech from the documents.

In the final tokenized corpus, I added the lowercase lemmas of the remaining tokens, so that different forms of the same word, such as “Language” and “languages”, result in the same token.

Afterwards, I used the Phrases collocation detection from Gensim, a statistical method that identifies words that generally co-occur, to build n-grams. By repeating the steps of training the model and running it on the corpus twice, I obtained bigrams, trigrams and even some 4-grams, which resulted from combining two bigrams. After adding the collocations to the corpus, I removed all of the unigrams, since around 70% of nouns or noun phrases that represent technical terms contain more than one word (Justeson & Katz, 1995).

# Detalii de implementare

În plus fata de capitolul precedent acesta conține elemente specifice ale rezolvării problemei care au presupus dificultăți deosebite din punct de vedere tehnic. Pot fi incluse configurații, secvențe de cod, pseudo-cod, implementări ale unor algoritmi, analize ale unor date, scripturi de testare. De asemenea, poate fi detaliat modul în care au fost utilizate tehnologiile introduse in capitolul 3.

Criterii pentru calificativul Nesatisfăcător:

* Sunt prezentate pe scurt scheme și pseudo-cod.

Criterii pentru calificativul Satisfăcător:

* Descriere sumara a implementării, prezentarea unor secvențe nerelevante de cod, scheme, etc.

Criterii pentru calificativul Bine:

* Descrierea detaliată a algoritmilor/structurilor utilizați; Prezentarea etapizată a dezvoltării, inclusiv cu dificultăți de implementare întâmpinate, soluții descoperite; (dacă este cazul) demonstrarea corectitudinii algoritmilor utilizați.

## Indicații formatare tabele

Se recomandă utilizarea tabelelor de forma celui de mai jos. Font: Calibri, 9.

Orice tabel prezent în teză va fi referit în text; exemplu: a se vedea Tabel 1.

Tabel 1 – Sumarizare criterii

|  |  |  |
| --- | --- | --- |
| Calificativ | Criteriu | Observații |
| Nesatisfăcător | Sunt prezentate pe scurt scheme și pseudo-cod |  |
| Satisfăcător | Descriere sumara a implementării, prezentarea unor secvențe nerelevante de cod, scheme, etc. |  |
| Bine | Descrierea detaliată a algoritmilor/structurilor utilizați; Prezentarea etapizată a dezvoltării, inclusiv cu dificultăți de implementare întâmpinate, soluții descoperite; (dacă este cazul) demonstrarea corectitudinii algoritmilor utilizați. | Pot fi incluse configurații, secvente de cod, pseudo-cod, implementări ale unor algoritmi, analize ale unor date, scripturi de testare. |

# Studiu de caz / Evaluarea rezultatelor

Acest capitol trebuie să răspundă, în principiu, la **2 întrebări** și să se încheie cu **o discuție** a rezultatelor obținute. Cele doua întrebări la care trebuie sa se răspundă sunt:

1) **Merge corect**? (Conform specificațiilor extrase în capitolul 2);

Evaluarea dacă merge corect se face pe baza cerințelor identificate în capitolele anterioare.

2) Cât de bine merge / cum se compară cu soluțiile existente? (pe baza unor metrici clare).

Evaluarea cât de bine merge trebuie să fie bazată pe procente, timpi, cantitate, numere, **comparativ cu soluțiile prezentate în capitolul 3**. Poate fi vorba de performanță, overhead, resurse consumate, scalabilitate etc.

În realizarea discuției, se vor utiliza tabele cu procente, rezultate numerice și grafice. În mod obișnuit, aici se fac comparații și teste comparative cu alte proiecte similare (dacă există) și se extrag puncte tari și puncte slabe. Se ține cont de avantajele menționate și se demonstrează viabilitatea abordării / aplicației, de dorit prin comparație cu alte abordări (dacă acest lucru este posibil). Cuvântul cheie la evaluare este „metrică“: trebuie să aveți noțiuni măsurabile și cuantificabile. În cadrul procesului de evaluare, explicați datele, tabelele și graficele pe care le prezentați și insistați pe relevanța lor, în următorul stil: „este de preferat ... deoarece …“; explicați cititorului nu doar datele ci și semnificația lor și cum sunt acestea interpretate. Din această interpretare trebuie să rezulte poziționarea proiectului vostru printre alternativele existente, precum și cum poate fi acesta îmbunătățit în continuare.

Criterii pentru calificativul *Nesatisfăcător*:

* Aplicația este testată dar rulează pe calculatorul studentului, nu există posibilități de testare, nu a fost validată cu clienți / utilizatori;
* Nu au fost realizate comparații cu alte sisteme similare.

Criterii pentru calificativul *Satisfăcător*:

* [Dezvoltare de produs] Există teste unitare și de integrare, există o strategie de punere în funcțiune (*deployment*), există validare minimală cu clienții / utilizatorii.
* [Cercetare] Principalele componente și soluția în ansamblu au fost evaluate din punct de vedere al performanței, însă nu sunt folosite seturi de date standard, există unele erori de interpretare a datelor.
* [Ambele] Discuție minimală asupra relevanței rezultatelor prezentate, comparație minimală cu alte sisteme similare.

Criterii pentru calificativul *Bine*:

* [Dezvoltare de produs] Teste unitare și de integrare, instrumente de punere in funcțiune (*deployment*) utilizate și care arată lucru constant de-a lungul semestrului, lucrare validată cu clienții / utilizatorii, produs în producție.
* [Cercetare] Componentele și soluția în ansamblu au fost evaluate din punct de vedere al performanței, folosind seturi de date standard și cu o interpretare corectă a rezultatelor.
* [Ambele] Discuție cu prezentarea calitativă și cantitativă a rezultatelor, precum și a relevanței acestor rezultate printr-o comparație complexă cu alte sisteme similare.

# Concluzii

În acest capitol este sumarizat întreg proiectul, de la obiective, la implementare, si la relevanta rezultatelor obținute. În finalul capitolului poate exista o subsecțiune de „Dezvoltări ulterioare“.  
Criterii pentru calificativul *Nesatisfăcător*:

* Concluziile nu sunt corelate cu conținutul lucrării;

Criterii pentru calificativul *Satisfăcător*:

* Concluziile sunt corelate cu conținutul lucrării, însă nu se oferă o imagine asupra calității și relevantei rezultatelor obținute;

Criterii pentru calificativul *Bine*:

* Concluziile sunt corelate cu conținutul lucrării, și se oferă o imagine precisa asupra relevantei și calității rezultatelor obținute în cadrul proiectului.

# REFERENCES

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131. https://doi.org/https://doi.org/10.1016/j.is.2022.102131

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). YAKE! Collection-Independent Automatic Keyword Extractor. *Information Retrieval Technology*, 806–810. https://doi.org/10.1007/978-3-319-76941-7\_80

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. https://doi.org/10.1016/J.INS.2019.09.013

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl\_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. *ArXiv Preprint ArXiv:1612.03651*.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *ArXiv Preprint ArXiv:1607.01759*.

Justeson, J., & Katz, S. (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, *1*, 9. https://doi.org/10.1017/S1351324900000048

Matthew Honnibal. (n.d.). *Dead Code Should Be Buried*. Retrieved June 9, 2024, from https://explosion.ai/blog/dead-code-should-be-buried

Moreno-Ortiz, A. (2024). *Keywords* (pp. 59–102). https://doi.org/10.1007/978-3-031-52719-7\_4

Nomoto, T. (2022). Keyword Extraction: A Modern Perspective. *SN Computer Science*, *4*. https://doi.org/10.1007/s42979-022-01481-7

Řeh uřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New for NLP Frameworks*, 45–50.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In *Text Mining: Applications and Theory* (pp. 1–20). https://doi.org/10.1002/9780470689646.ch1

*spaCy*. (n.d.). Retrieved June 9, 2024, from https://spacy.io/

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Tijare, P., & Rani, P. (2020). Exploring popular topic models. *Journal of Physics: Conference Series*, *1706*, 12171. https://doi.org/10.1088/1742-6596/1706/1/012171

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*(3), 1169–1180.

# Anexe

1. https://demos.explosion.ai/displacy [↑](#footnote-ref-1)
2. https://demos.explosion.ai/displacy-ent [↑](#footnote-ref-2)
3. https://fasttext.cc/docs/en/language-identification.html [↑](#footnote-ref-3)
4. https://radimrehurek.com/gensim/index.html [↑](#footnote-ref-4)
5. https://radimrehurek.com/gensim/intro.html [↑](#footnote-ref-5)
6. https://radimrehurek.com/gensim/models/ldamulticore.html [↑](#footnote-ref-6)
7. https://pypi.org/project/yake/ [↑](#footnote-ref-7)