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](#_Toc7074723)

[Abstract 2](#_Toc7074724)

[Mulțumiri 3](#_Toc7074725)

[1 Introducere 4](#_Toc7074726)

[1.1 Context 4](#_Toc7074727)

[1.2 Problema 4](#_Toc7074728)

[1.3 Obiective 4](#_Toc7074729)

[1.4 Structura lucrării 4](#_Toc7074730)

[2 Analiza și specificarea cerințelor 5](#_Toc7074731)

[3 Studiu de piață / Abordări existente 6](#_Toc7074732)

[3.1 Indicații formatare figuri 6](#_Toc7074733)

[4 Soluția propusă 8](#_Toc7074734)

[4.1 Indicații formatare formule 8](#_Toc7074735)

[5 Detalii de implementare 9](#_Toc7074736)

[5.1 Indicații formatare tabele 9](#_Toc7074737)

[6 Studiu de caz / Evaluarea rezultatelor 11](#_Toc7074738)

[7 Concluzii 13](#_Toc7074739)

[8 Bibliografie 14](#_Toc7074740)

[9 Anexe 16](#_Toc7074741)

# 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 keywords and topics 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 accurate 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 and fastText for language detection.

## 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 texts from domains that it was not trained on.

Despite being an unsupervised method, TFIDF (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 in all of the documents. 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 to pursue in my implementation. YAKE ranks candidate keywords using a score calculated 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 stopwords 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, stopwords 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 starting with a 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 term, 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 stopwords, 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), with 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 stopword.

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

# Soluția propusă

Capitolul conține o privire de ansamblu a soluției ce rezolvă problema, prin prezentarea structurii / arhitecturii acesteia. În funcție de tipul lucrării acest capitol poate conține diagrame (clase, distribuție, workflow, entitate-relație), demonstrații de corectitudine pentru algoritmii propuși de autor, abordări teoretice (modelare matematică), structura hardware, arhitectura aplicației.

Criterii pentru calificativul *Nesatisfăcător*:

* Descriere în limbaj natural.

Criterii pentru calificativul *Satisfăcător*:

* Descriere + diagrame de baze de date, workflow, clase, algoritmi.

Criterii pentru calificativul *Bine*:

* Descriere + diagrame de baze de date, workflow, clase, algoritmi + descrierea unui proces prin care s-a realizat arhitectura/structura soluției.

## Indicații formatare formule

Formulele matematice utilizate în document vor fi centrate în pagină și numerotate. Se vor utiliza fontul Cambria Math, de dimensiune 11. Pentru a insera o nouă ecuație, utilizați Insert > Quick Parts > AutoText > Ecuație.

Toate formulele prezente în lucrare vor fi referite în text. Exemplu: *Utilizând sistemul de Insert > Bookmark*, respectiv *Reference > Cross-reference* putem cita ecuația (1) respectiv ecuația (2), citările fiind actualizate și în urma unor adăugări/ ștergeri de ecuații, cu *Select All – Update Field*. Pentru mai multe detalii despre utilizarea acestui sistem de referire și formatare puteți consulta:

<https://www.youtube.com/watch?v=9YGTH4WrY_8>.

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  | () |

# 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

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

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

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

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

# Anexe