



Text Analytics in Bulgarian

SOFIA
UNIVERSITY



ST. KLIMENT
OHRIDSKI
1888

FACULTY
OF ECONOMICS
AND BUSINESS
ADMINISTRATION

What you will learn

- ◎ More about my PhD thesis in the field. 😊
- ◎ What is specific about the Bulgarian language? 🌸
- ◎ Language resources for Bulgarian? ⚙️
- ◎ Availability of language resources (for Bulgarian) in Python. 🚀

PhD thesis (1)

- ◎ **FEBA** – a place where the **industry** and the **academy** meet.
- ◎ **Industry-funded PhD** – work on a real business case as part of your dissertation.



PhD thesis (2)

- ⦿ Analyzed type of data – **chat data in Bulgarian** (communication between clients and operators in the bank's contact center):


```
2020-01-22.text - Notepad
File Edit Format View Help
Timestamp: 2020-01-22T15:56:19Z
Unread:false
Visitor ID: 1111111.aaaa1aaaaa11a
Visitor Name: Visitor 11111111
Visitor Email:
Visitor Notes:
IP: 111.11.111.1
Country Code: BG
Country Name: Bulgaria
Region:
City:
User Agent: Mozilla/1.1 (Windows NT 1.1) AppleWebKit/111.11 (KHTML, like Gecko) Chrome/11.1.1111.11 Safari/111.11
Platform: Windows
Browser: Chrome

(2020-01-22 15:56:19) Visitor 11111111: Здравейте! Трябва ми помощ - забравил съм потребителското си име и паролата за онлайн банкиране.
(2020-01-22 15:56:34) Иван Иванов: Здравейте!
(2020-01-22 15:57:08) Иван Иванов: Необходимо е да посетите офис на банката, за да получите нови потребителско име и парола.
(2020-01-22 15:57:08) Visitor 11111111: Много ви благодаря!
(2020-01-22 15:57:08) Visitor 11111111: Хубава вечер!
(2020-01-22 15:57:08) Иван Иванов: Моля, хубава вечер и на вас!
=====
```



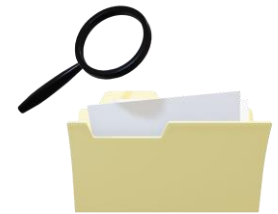
PhD thesis (3)

Chats between clients and operators are a **valuable data source** which could be used in order to:

- ◎ Find the **main topics of client interest**;
 - ◎ Inspect **customer satisfaction** (sentiment) with the services;
 - ◎ **Develop training data** for a chatbot/dialogue system.
- 

Text Analytics in Bulgarian: An Overview

- © What has been accomplished by the researchers in the field of NLP and text analytics in Bulgarian?
- © Main areas of interest:
 - **Available language resources**
 - **Research with practical applications**
- © Hristova, Gloria. "**Text Analytics in Bulgarian: An Overview and Future Directions.**" Cybernetics and Information Technologies 21, no. 3 (2021): 3-23.



Language resources/Tools for text data in Bulgarian (1)

© The **BulTreeBank** group -

<https://bultreebank.org/bg/resources/>

- **Stop words list** for Bulgarian
- The **BulTreeBank corpus** -
<http://bultreebank.org/en/resources/short-description-morphologically-annotated-part-bultreebank-bultreebank-morph/>
- **POS tagger** for Bulgarian – available in Python! -
<https://www.cis.lmu.de/~schmid/tools/TreeTagger/>
<https://pypi.org/project/treetaggerwrapper/>

Language resources/Tools for text data in Bulgarian (2)

- ◎ **Stemming tool** developed by P. Nakov - <https://pypi.org/project/bulstem/>
- ◎ Several **text processing tools** for Bulgarian available in Python – <https://pypi.org/project/classla/>
- ◎ **More information** on available language resources for Bulgarian:
<http://www.clarin.si/info/k-centre/faq4bulgarian/>
<https://lremap.elra.info/?languages=Bulgarian>
- ◎ Check out the **resources in Hugging Face** - <https://huggingface.co/models?language=bg&sort=downloads>

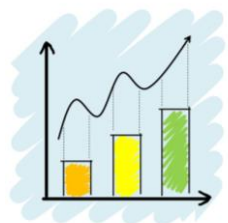
Practical applications of text analytics in Bulgarian

- ◎ **Fake news** detection;
- ◎ **Toxic/misleading behavior** in community forums;
- ◎ A lot of progress in the **biomedical NLP domain** – information extraction from patient records; detection of risk factors and diseases; discovering relations between the treatment of a disease and other disorders etc.



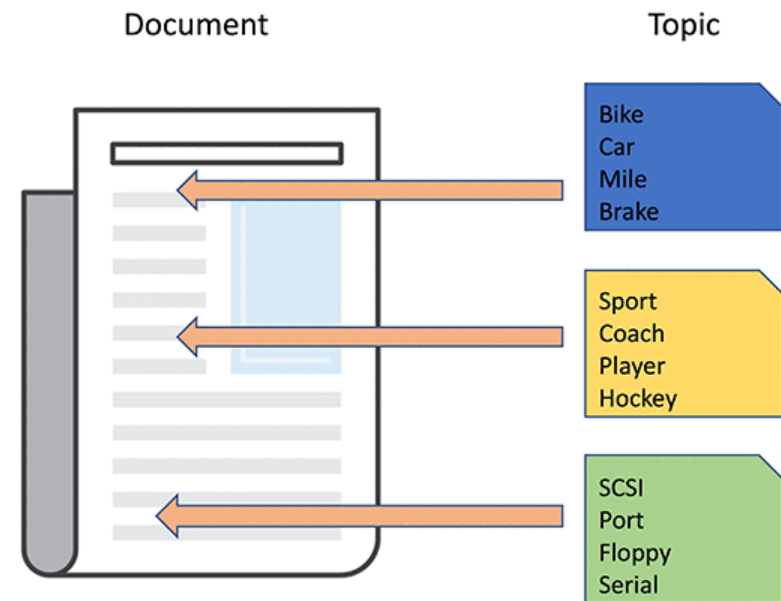
What is different about the Bulgarian language compared to other languages (from practical point of view)?

- ◎ Bulgarian is a “**low-resource language**” – only a few linguistic resources and technologies for working with text data in Bulgarian have been developed.
- ◎ Examples of **high-resource languages** – English, Chinese, Spanish.
- ◎ The extent to which a language falls into one category or another depends mainly on **economic, social, and cultural factors**.
- ◎ **Purely statistical techniques can be applied on textual data in any language!**



Topic modeling – know your customer

- ◎ Topic modeling – a statistical technique for **discovering the abstract "topics"** that occur in a collection of documents.
- ◎ Topics consist of “**related keywords**”.
- ◎ **Unsupervised technique** for discovery of hidden semantic structures in text data.



Predict customer sentiments

- ◎ Analytical problem:

Predict customer service rating based only on dialogue (chat) data.

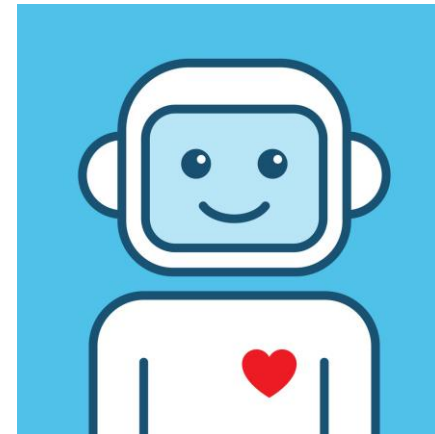
- ◎ Customers may evaluate chat communication either **positively** or **negatively**.

- ◎ Sentiment analysis?



Extract similar customer requests

- © Why?
- © Application in chatbot development - similar customer requests (for example, frequently asked questions) might serve as training data.
- © A search engine for client requests...
- © Word embeddings?





Thanks!

Any questions?

You can find me at:
`g.hristova@feb.uni-sofia.bg`

