University *of Ljubljana*
Faculty *of Computer and Information Science*

IMapBook collaborative discussions classification

Marko Katrašnik and Erica Zago

Abstract

- First defense (10 points): task selection simple corpusprocessing/analysis
  - Introduction, existing solutions, initial ideas.
- Interim defense (10 points): at least one example of a solution to aproblem
  - Introduction, related work, implemented baseline, future directions
- Final defense (30 points): full submission and presentation
  - Clean Git repository (fully reproducible) and final report

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

KAJ DELAMO? Napovedati moramo CodePreliminary
-what is the purpose of the project, what will we solve -what solutions for the problem already exist
V tem projektu se bomo ukvarjali s klasifikacijo oziroma napovedovanjem danih sporočil. Sporočila so bila napisana v okviru raznih knjižnih krožkov, kjer so različne skupine sodelovale. Vsaka skupina je imela nalogo, da skupaj sestavi odgovor na podano vprašanje, nanašajoč se na preprano knjigo/poglavje. Preko sporočil so se zmenili kako odgovorit. Ta sporočila so bila tako različnih vrst, kot npr. komentarji, diskusije, vprašanja, navodila itd. Naš cilj je klasificirati vsako sporočilo v pripadajočo kategorijo.
V tem delu se bomo ukvarjali s klasifikacijo govornih dejanj (angl. speech act classification), kjer bi besedilo najprej prečistili in nato poskusili različne metode napovedovanja klasifikacije.

Analiza podatkov

Število komentarjev/sporočil, ki moramo napovest

- 711 sporočil v prvem tab-u
- 130 sporočil v drugem tab-u

V tabeli 1 so prikazane kategorije spročil v zavihku CREW data, kjer imajo sporočila tudi pripadajoči skupni končni odgovor. V tabeli 2 pa so prikazane kategorije iz zavihka Discussion, kjer ni pripadajočega skupnega odgovora.

Opazna je zelo neenakomerna porezdelitev števila soročil po kategorijah, kar bo potrebno upoštevati pri uporabi algoritmov, še bolj pa pri uporabi in interpretaciji metrik za evalvacijo.

Na slikah 1 in 2 sta prikazani distribuciji dolžin sporočil (število besed v sporočilu) v zavihkih Crew data in discussion. Distribuciji sta si precej različni, v prvi je veliko več kratkih sporočil, druga je bolj enakomerna, prisotnih je tudi nekaj daljših sporočil. Povprečna dolžina sporočila v CREW data je 11 besed, v discussion pa 42,3 besed.

Methods

Postopek klasifikacije sporočil

- Predprocesiranje in vektoriziranje podatkov
- Izbira modela
- Izbira parametrov, s katerimi bo model delal

Predprocesiranje in vektoriziranje

Predprocesiranje podatkov:

- Lower-case the text

Table 1. CREW data - Razredi in število sporočil v vsakem razredu.

Kategorija	št. sporočil
assignment instructions	36
assignment instructions question	1
assignment question	10
content discussion	299
content question	18
discussion wrap-up	12
emoticon/non-verbal	12
external material	2
feedback	31
general comment	55
general discussion	9
general question	2
greeting	11
incomplete/typo	10
instruction question	3
logistics	80
observation	2
opening statement	8
outside material	10
response	99

Table 2. Discussion - Razredi in število sporočil v vsakem razredu.

Kategorija	št. sporočil
assignment instructions	1
content discussion	94
content question	5
emoticon/non-verbal	3
feedback	3
general comment	6
incomplete/typo	10
instruction question	1
response	7

- Remove all non-ASCII characters and punctuation
- Convert white-space blocks to single spaces
- Stem words
- Remove common (e.g. stop) words
- v članku kang2013 predlaga POS-tag

Uporabili bi N-gram vektorje, podrobneje unigrame in bigrame. Vektorji bi seveda bili numerični in za pridobitev teh bi uporabili TF-IDF metriko. TF-IDF je vrsta meritve, ki ocenjuje, koliko je pomembna neka beseda za dokument v zbirki dokumentov. Izračuna se tako, da pomnožimo dve metriki: kolikokrat se beseda pojavi v dokumentu (angl. Term Frequency) in obratno pogostost dokumenta v nizu dokumentov

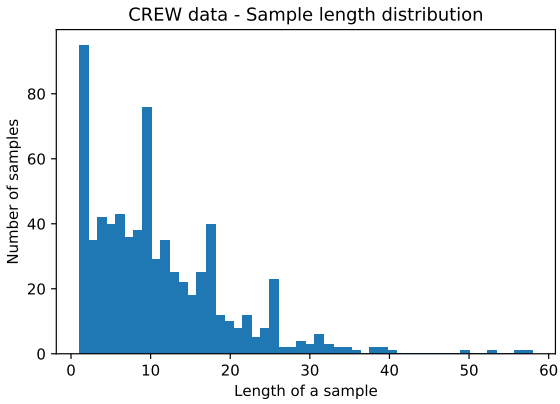

Figure 1. Distribucija dolžine sporočil v zavihku CREW data.

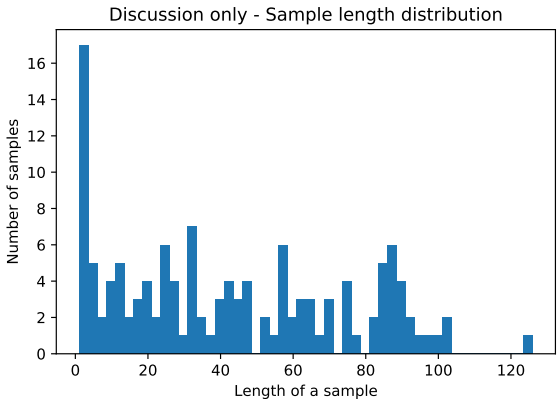

Figure 2. Distribucija dolžine sporočil v zavihku discussion.

(angl.Inverse Document Frequency). Tokenizacija - zaporedje besed

Izbira modela

Izbrana modela za klasifikacijo bosta MLP in BERT, katera ju boo primerjala in izluščila glavne prednosti.

MLP (angl. Multilayer Perceptrons) je klasični tip nevronske mreže. Sestavljen je iz ene ali več plasti nevronov. Podatki so vnešeni v vhodno plast, nato pa gredo skozi eno ali več skritih plasti, ki zagotavljajo ravni abstrakcije. Na izhodni plasti se pa naredijo napovedi. MLP je primeren za klasifikacijso napovedovanje, kjer je vhodnim podatkom dodeljen razred ali oznaka.

BERT (angl. Bidirectional Encoder Representations from Transformers) je dvosmerni transformator, ki je bil predhodno treniran s kombinacijo modeliranja z zamaskiranim jezikom in napovedovanja naslednjega stavka na velikem korpusu, ki ga sestavljata Toronto Book Corpus in Wikipedia.

Izbira parametrov, s katerimi bo model delal

Najpomembnejši parameter bo ”Message”, kateri vsebuje samo sporočilo, kateremu želimo napovedati kategorijo.

isAnswer in Page lahko zanemarimo.

Kategorijo ”Content Discussion” napovemo lahko z uporabo parametra ”Collab Response” če je podobno

Topic verjetno ne bo koristno.

Dobili bi lahko ključno besedo v kategorijah

Uporabni podatki: Message Time, Collab Response, povezava z besedilom knjige in predhodno sporočilo

Mogoče uporabno Pseudonym, da vemo če je več zaporednih sporočil pisala ena sama oseba.

Uporaba besedila knjig in skupnega končnega odgovora

Želiva ugotoviti tudi, ali ima uporaba podatkov besedila knjig in skupnega končnega odgovora pozitiven vpliv na uspešnost napovedovanja razredov posameznih sporočil. Glede na to, da ena podmnožica sporočil (discussion) nima podanega podatka o končnem odgovoru, ima pa precej daljša besedila sporočil, bova lahko primerjala tudi koristnost teh dveh lastnosti.

Besedila knjig in skupne končne odgovore bi uporabila tako, da bi izračunala podobnost s sporočilom. Predvidevava namreč, da so si sporočila, v katerih so govorili o vsebini (npr. content discussion) bolj podobna besedilu knjige in končnemu odgovoru. Tako kot pri osnovni klasifikaciji sprocil, bi tudi tukaj najprej za vektorizacijo besedil uporabila tf-idf, kasneje pa bi ga nadomestila z word2vec. Podobnost bi računala s kosinusno podobnostjo.

Podatka o podobnosti z besedilom knjige in končnim odgovorom, bi nato skupaj z napovedjo razreda sporočila (ta napoved bi bila pridobljena z zgoraj opisanimi postopki, ki uporabljajo samo vsebno sporočila) poslala na vhod še enega algoritma strojnega učenja, uporabila bi random forest, kasneje pa morda še MLP.

Okvirna shema algoritma klasifikacije sporočil z uporabo besedil knjig in končnih skupnih odgovorov je prikazana na sliki 3 .

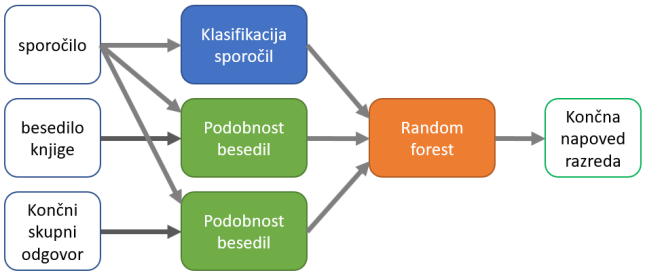

Figure 3. Okvirna shema uporabe besedila knjig in končnih odgovorov.

Drugo

Koristni linki (Področja/sorodna dela - Existing solutions)

- speech act analysis
- https://huggingface.co/transformers/
- https://developers.google.com/machine-learning/guides/text-classification/step-2-5
- https://cs224d.stanford.edu/reports/AbajianAaron.pdf
- Classification of Online Discussions Via Content and Participation (str 820-828): https://link.springer.com/content/pdf/10.1007%2F11881223.pdf
- https://www.sciencedirect.com/science/article/abs/pii/S016786551300 Nekaj razlaga o uporabi naivnega Bayesa: http://www.cs.memphis.edu/~vrus/publications/2010/Arabic-SAC.LubnaRusGraesser.pdf
- Tu so predlagani (poleg SVM in naiv B) tudi Random forest in k-nearest neighborshttps://arxiv.org/ftp/arxiv/papers/1901/1901.03904.pdf
- Poleg ostalih je predlagana Logistic regression https://arxiv.org/pdf/1605.05156.pdf

Initial ideas

- Kaj in kako lahko analiziramo: Napoved je odvisna od značilk - te vplivajo na napoved
  - Na podlagi vsebine komentarjev in teksta knjige, lahko napovemo na katero knjigo se nanašajo komentarji –¿ za kontent discussion ; nova spremenljivka za podobnost konentarja in vsebini knjige
  - Na podlagi komentarjev in CodePreliminary v train setu, napovemo test
  - Na podlagi prejšnjega komentarja (npr. če je to vprašanje) lahko predpostavimo da bo naslednji komentar odgovor?
  - lahko uporabimo Pseudonym, da vemo kdo govori in če je ta postavil set komentarjev (en komentar za drugim)
  - Uporaba slovarjev - npr. če vemo da je omejeno na neke znake ali na neke besede (npr. emoticons, greeting..., verjetno če je na koncu vprašaj bo vprašanje...)
  - Ker je dosti kategorij, jih bomo nekaj združili nadkategorije
  - Glede uporabe besedil knjig in končnih skupnih odgovorov bi predlagal tako: Za računanje podobnosti - Najprej uprabiva td-idf in kosinusno razdaljo, kasneje pa td-idf zamenjava za kakšen word embedding, npr word2vec, ali bert, če se izkaže da je uporaben tako
- Profesor predlagal:
  - TF-IDF
  - SVM
  - NEVRONSKE Mreže
- Testiranje napovedovanja:

  - cross validation
  - (profesor je omenil, da jih ne smemo random vzet in da moramo pazit na predhodna poročila)
  - set podatkov razdelimo na dva dela (train in test set) v razmerju 80% - 20% (prvi tab: 568 - 143; drugi tab: 104 - 26)
  - train set se še enkrat razdeli (v train in validation set) v razmerjo 80% - 20% (prvi tab: 454 - 114; drugi tab: 83 - 21)
    https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

- Ocenjevanje izvedene metode:
  - precision and recall
  - Ocena f (poroča o uspešnosti):
    * utežena
    * mikro in makro
    * ker imamo več razredov sproti poročamo povprečje
    * so implementirane v gitlearnu

Equations

You can write equations inline, e.g. $\cos\pi = -1$, $E = m \cdot c^2$ and α, or you can include them as separate objects. The Bayes's rule is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are some events. You can also reference it – the equation 1 describes the Bayes's rule.

Lists

We can insert numbered and bullet lists:

1. First item in the list.
2. Second item in the list.
3. Third item in the list.

- First item in the list.
- Second item in the list.
- Third item in the list.

We can use the description environment to define or describe key terms and phrases.

Word What is a word?.

Concept What is a concept?

Idea What is an idea?

Random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Figures

You can insert figures that span over the whole page, or over just a single column. The first one, Figure 4, is an example of a figure that spans only across one of the two columns in the report.

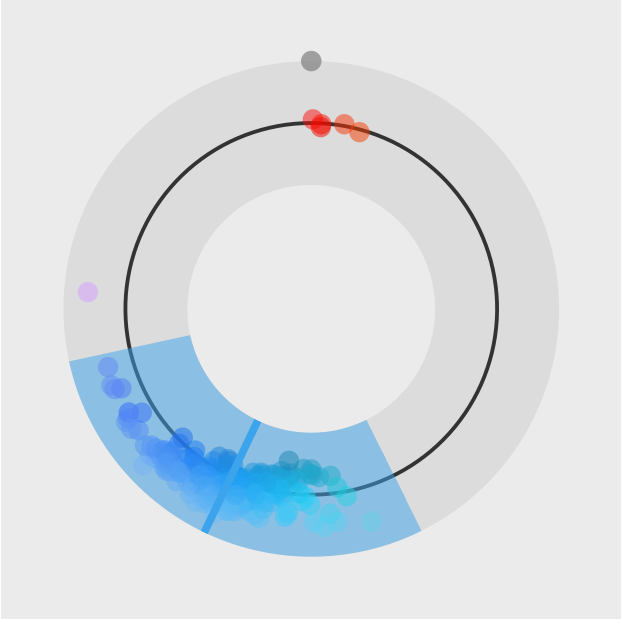

Figure 4. A random visualization. This is an example of a figure that spans only across one of the two columns.

On the other hand, Figure 5 is an example of a figure that

spans across the whole page (across both columns) of the report.

Tables

Use the table environment to insert tables.

Table 3. Table of grades.

Name		
First name	Last Name	Grade
John	Doe	7.5
Jane	Doe	10
Mike	Smith	8

Code examples

You can also insert short code examples. You can specify them manually, or insert a whole file with code. Please avoid inserting long code snippets, advisors will have access to your repositories and can take a look at your code there. If necessary, you can use this technique to insert code (or pseudo code) of short algorithms that are crucial for the understanding of the manuscript.

Listing 1. Insert code directly from a file.

```
import os
import time
import random

fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
```

Listing 2. Write the code you want to insert.

```
import(dplyr)
import(ggplot)

ggplot(diamonds,
           aes(x=carat, y=price, color=cut)) +
  geom_point() +
  geom_smooth()
```

Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

More random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Nulla rhoncus tortor eget ipsum commodo lacinia sit amet eu urna. Cras maximus leo mauris, ac congue eros sollicitudin ac. Integer vel erat varius, scelerisque orci eu, tristique purus. Proin id leo quis ante pharetra suscipit et non magna. Morbi in volutpat erat. Vivamus sit amet libero eu lacus pulvinar pharetra sed at felis. Vivamus non nibh a orci viverra rhoncus sit amet ullamcorper sem. Ut nec tempor dui. Aliquam convallis vitae nisi ac volutpat. Nam accumsan, erat eget faucibus commodo, ligula dui cursus nisi, at laoreet odio augue id eros. Curabitur quis tellus eget nunc ornare auctor.

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

Acknowledgments

Here you can thank other persons (advisors, colleagues ...) that contributed to the successful completion of your project.

References

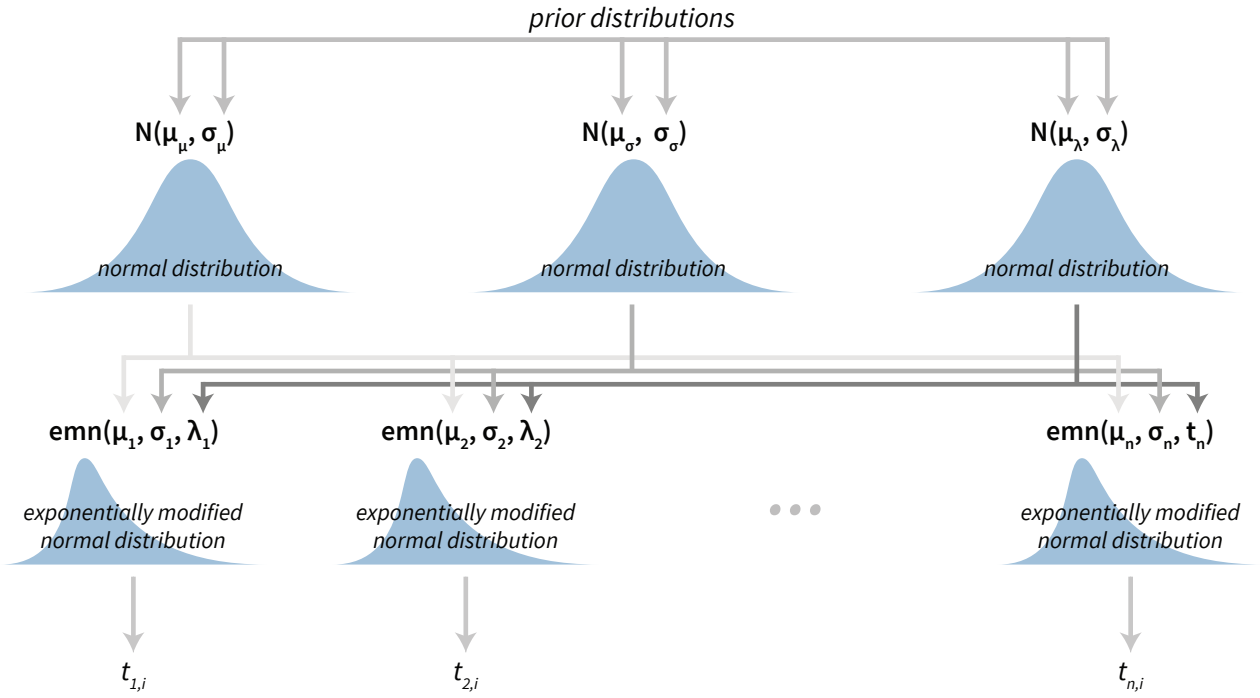

Figure 5. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.