

Sistemi za istraživanje i analizu podataka

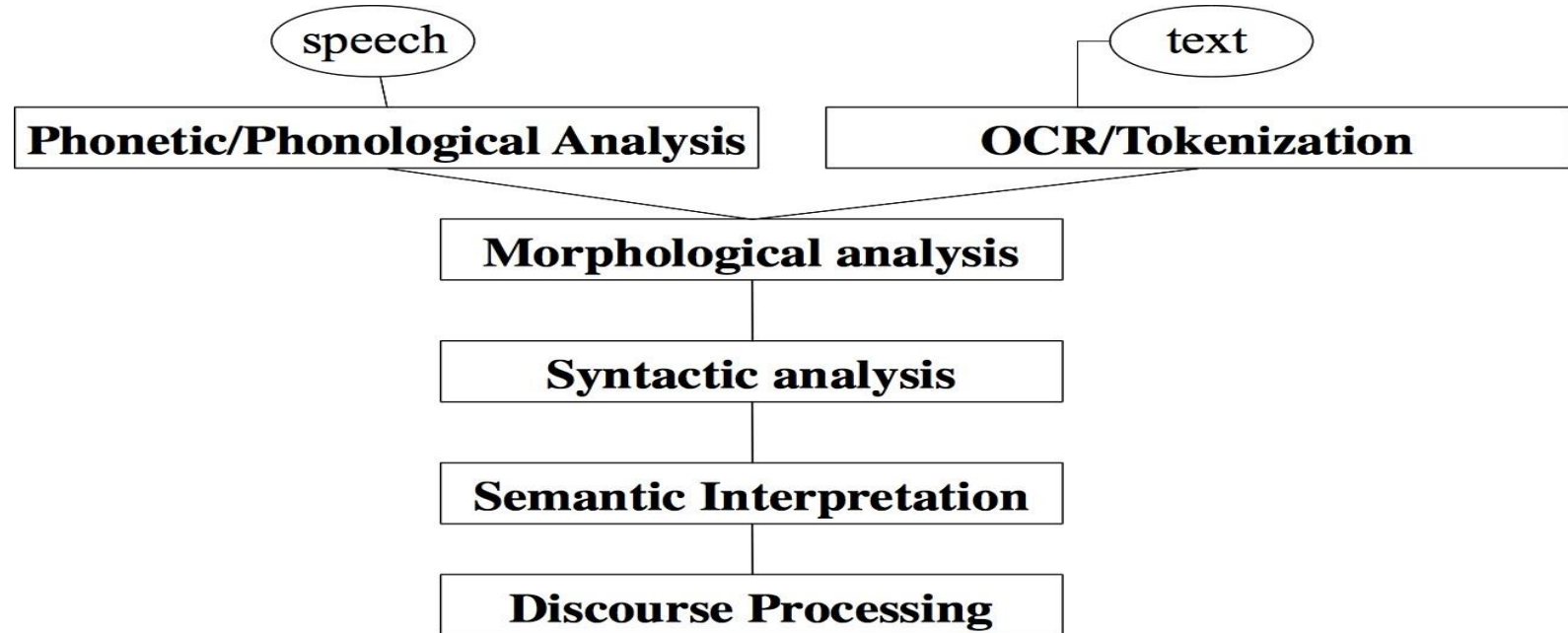
NLP

predavač: Aleksandar Kovačević

Šta je NLP?

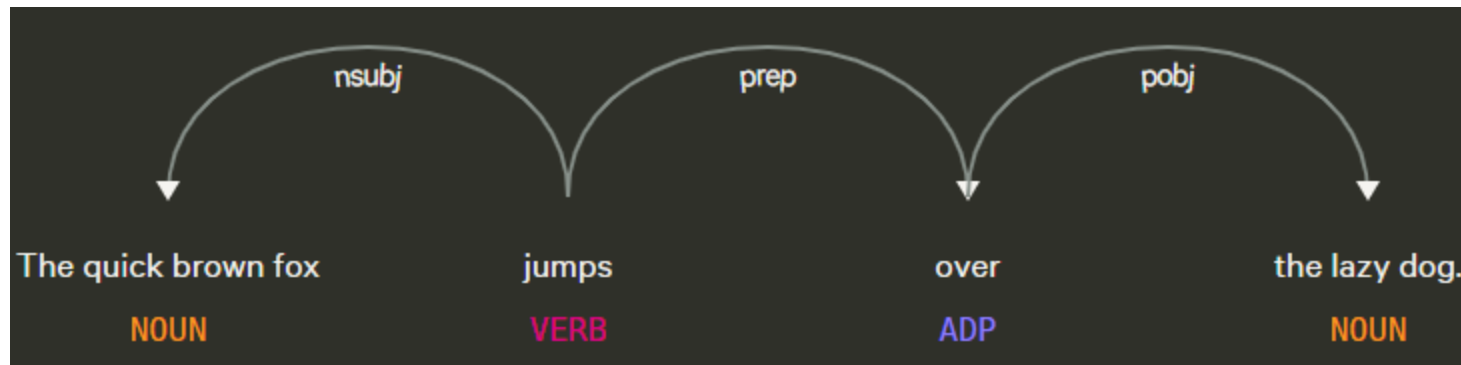
- NLP je oblast koja je u preseku tri velike oblasti:
 - računarstva
 - veštačke inteligencije
 - i lingvistike
- Cilj: naučiti računare da procesiraju i “razumeju” prirodan jezik tako da mogu da obavljaju korisne zadatke kao na primer:
 - određivanje teme ili sentimenta teksta, prepoznavanje entiteta u tekstu
 - odgovaranje na pitanja i obavljanje zadataka
 - Digitalni asistenti (zakazivanje sastanaka, internet kupovina...), npr. Siri, Google Assistant, Facebook M, Cortana ...
 - mašinski prevod
- Potpuno razumevanje jezika od strane računara je težak cilj, ali savršen za oblast veštačke inteligencije.

Nivoi funkcionisanja NLP



Sintaksna analiza

- Sintaksa se bavi gramatičkom strukturom rečenice.
- U sledećem primeru rečenica ima glagolsku frazu i imeničku frazu, pored toga svaka reč ima označenu svoju vrstu (imenica, zamenica itd.)

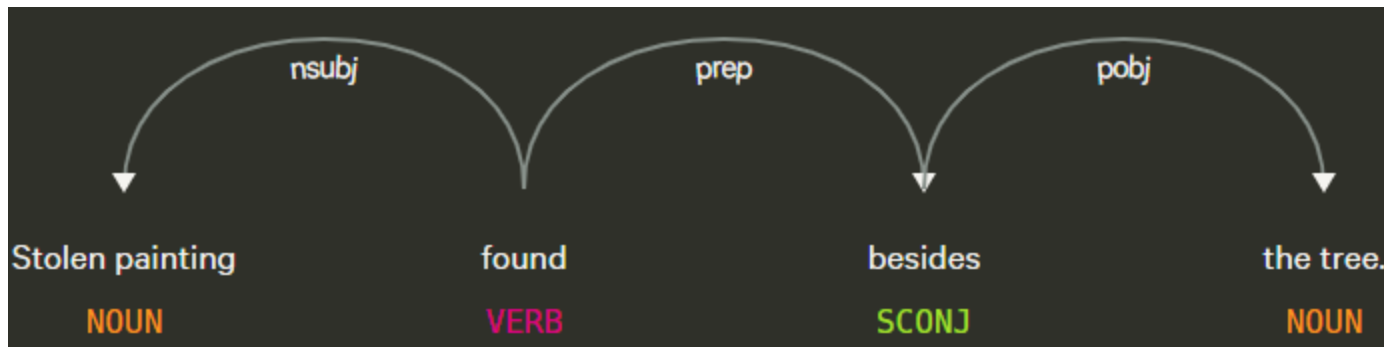


https://explosion.ai/demos/displacy?text=The%20quick%20brown%20fox%20jumps%20over%20the%20lazy%20dog.&model=en_core_web_sm&cpu=1&cph=1

- Iako su rezultati sintaksne analize veoma korisni za kompleksnije NLP zadatke, sintaksni parseri nisu na dovoljnom nivou za razumevanje jezika.

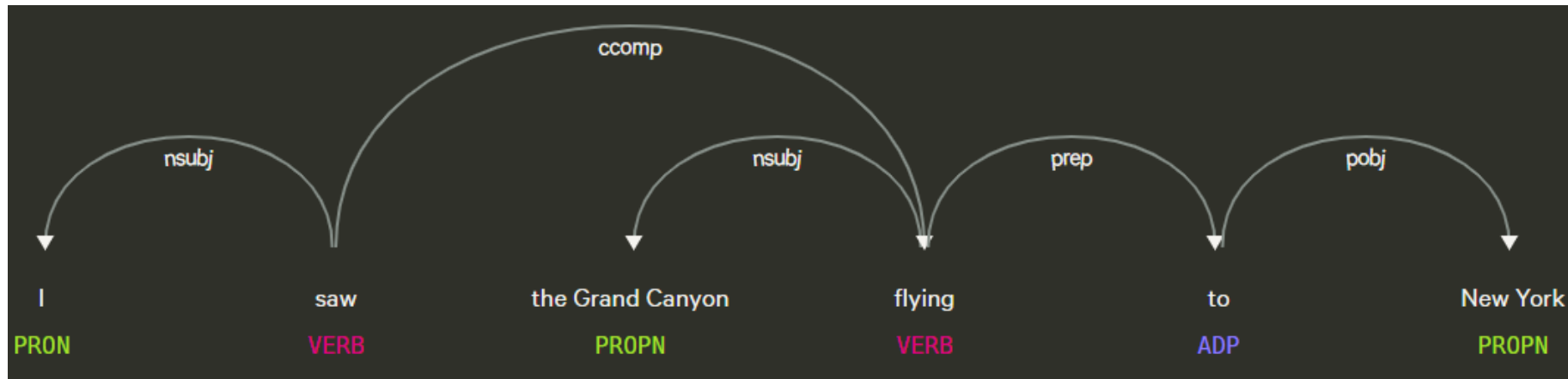
Sintaksna analiza je teška za računare

- Obe rečenice imaju isto značenje za nas, ali ne i za računar.

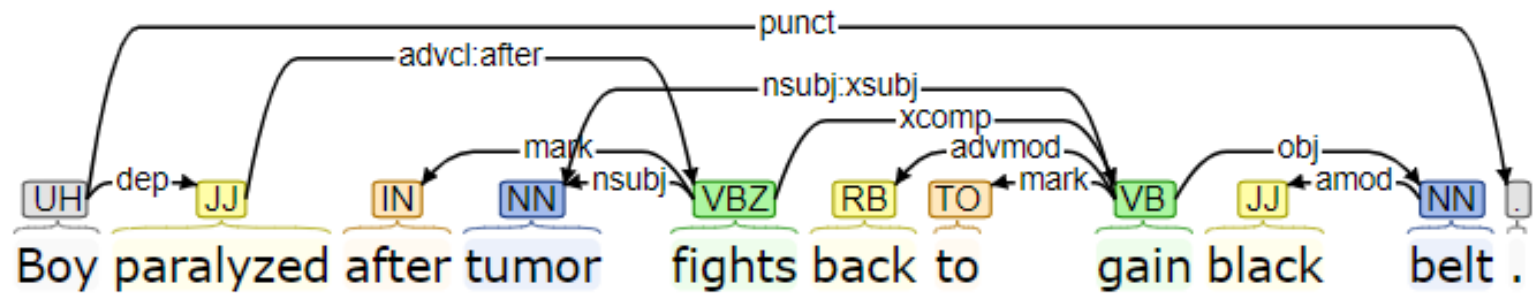


Sintaksna analiza je teška za računare

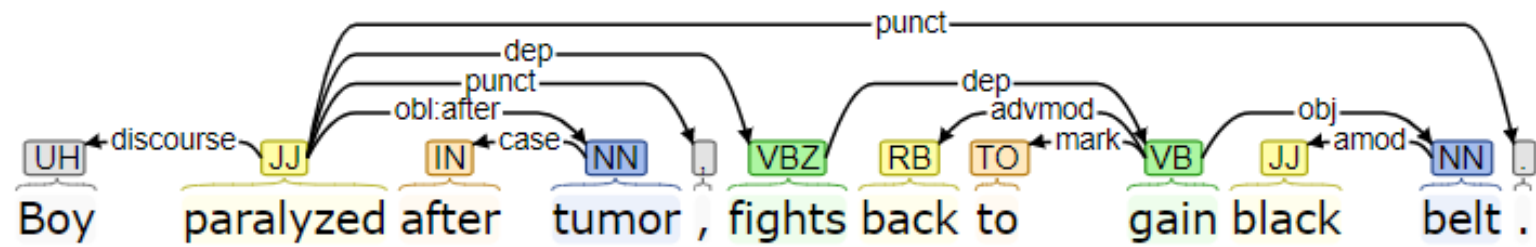
- Evo par realnih primera koji su teški za prasere.



Sintaksna analiza je teška za računare



<https://corenlp.run/>



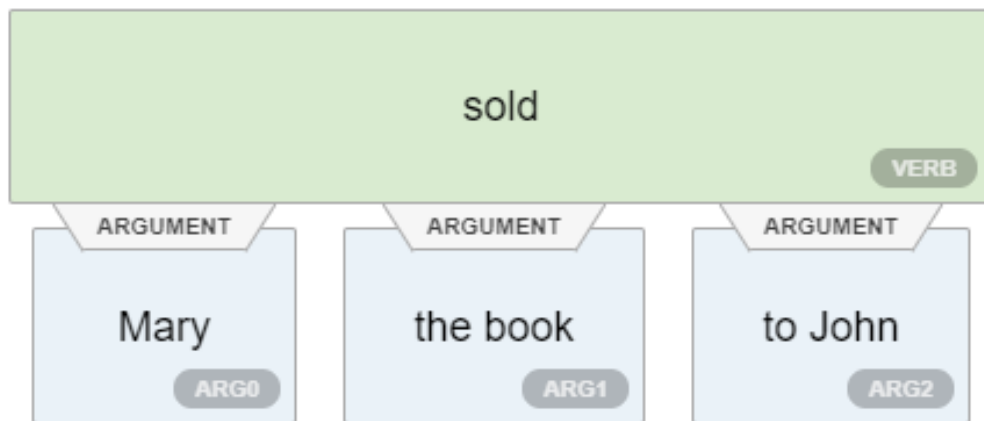
- Interpunkcija pomaže, ali ljudi razumeju obe rečenice, a cilj nam je da računari postignu isti nivo razumevanja.

Semantička analiza

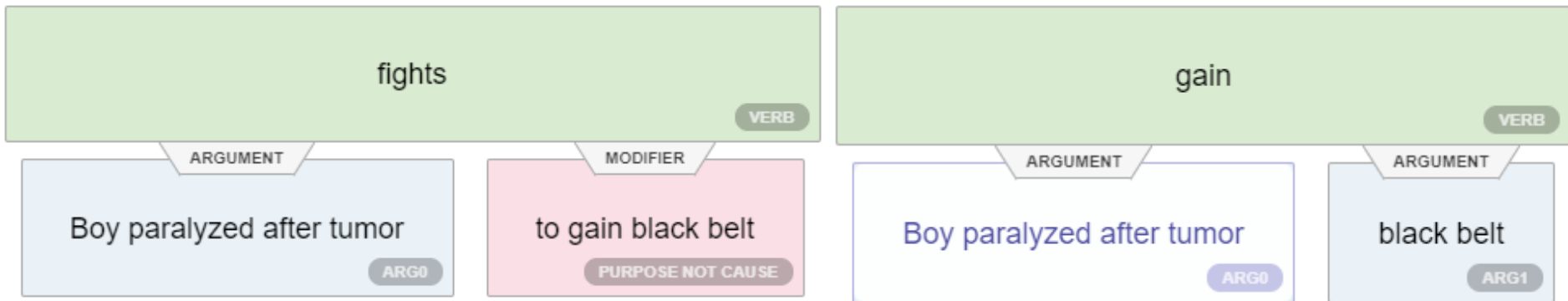
- Otkrivanje značenja pre svega reči u rečenici, pa onda cele rečenice ili celog dokumenta.
- Veoma težak zadatak za računare.
- Poznate tradicionalne metode:
 - Dodela semantičkih uloga (*Semantic role labeling, SRL*)
 - Semantičko parsiranje

Semantic role labeling

- Jedan od ranijih metoda (koji se idalje razvija).
- Ideja je da se delovima rečenice dodele uloge koje reprezentuju njihovu semantiku.



SRL može da radi jako dobro



<https://demo.allennlp.org/semantic-role-labeling/MjUwMjA5OQ==>

Semantičko parsiranje

- Postupak prevođenja teksta u neki od mašinskih jezika.

“There is at least one black block on a blue block.”



> @start@ -> bool

> bool -> [<Set[Object],int:bool>, Set[Object], int]

> <Set[Object],int:bool> -> object_count_greater_equals

> Set[Object] -> [<Set[Object]:Set[Object]>, Set[Object]]

> <Set[Object]:Set[Object]> -> black

> Set[Object] -> [<Set[Object]:Set[Object]>, Set[Object]]

> <Set[Object]:Set[Object]> -> above

> Set[Object] -> [<Set[Object]:Set[Object]>, Set[Object]]

> <Set[Object]:Set[Object]> -> blue

> Set[Object] -> all_objects

> int -> 1

Semantička analiza u 2021 godini

- Iako se i dalje radi na razvoju semantičkog parsiranja, SRL i sličnih metoda,
- većina naučne i industrijske zajednice usmerila se u drugom pravcu određivanja semantike prirodnog jezika.
- Taj drugi pravac je obučavanje ogromih (~170 milijardi težina) dubokih neuronskih mreža (transformera),
- koji semantiku reči reprezentuju pomoću vektora.

Semantička analiza u 2021 - SuperGLUE

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

Alex Wang*
New York University

Yada Pruksachatkun*
New York University

Nikita Nangia*
New York University

Amanpreet Singh*
Facebook AI Research

Julian Michael
University of Washington

Felix Hill
DeepMind

Omer Levy
Facebook AI Research

Samuel R. Bowman
New York University

Abstract

In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. The GLUE benchmark, introduced a little over one year ago, offers a single-number metric that summarizes progress on a diverse set of such tasks,

SuperGLUE - tasks

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 _a /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

SuperGLUE - tasks

BoolQ **Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*

Question: *is barq's root beer a pepsi product* **Answer:** No

CB **Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?*

Hypothesis: *they are setting a trend* **Entailment:** Unknown

COPA **Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*
Alternative 1: *The sun was rising.* **Alternative 2:** *The grass was cut.*
Correct Alternative: 1

MultiRC **Paragraph:** *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*

Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)*

SuperGLUE - tasks

ReCoRD **Paragraph:** *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*

Query For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency **Correct Entities:** US

RTE **Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*





Hypothesis: *Christopher Reeve had an accident.* **Entailment:** False

WiC **Context 1:** *Room and board.* **Context 2:** *He nailed boards across the windows.*
Sense match: False

WSC **Text:** *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** False

SuperGLUE – najbolji timovi - 2020

<https://super.gluebenchmark.com/leaderboard>











Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
3	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
4	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
5	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
6	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
7	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1

SuperGLUE – najbolji timovi - 2021

<https://super.gluebenchmark.com/leaderboard>

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
2	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
4	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
5	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
6	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
7	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2

SuperGLUE – najbolji timovi - 2023

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
10	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6

Model	LAMBADA Perplexity	SuperGLUE Score
GPT-2	8.4	44.5
GPT-3	3.8	71.8
GPT-4	2.6	89.5

SuperGLUE – šta nam rezultati govore?

- Veliki modeli (RoBERTa - 330M, T5 - 11B...).
- Svi su varijacije BERT transformera dobijene pre svega:
 - Dodavanjem više slojeva
 - Obučavanjem na većem korpusu (120 milijardi reči u T5 korpusu)
- Šta možemo da zaključimo?
- Bolji hardver + veći korpus = bolji rezultati. Granice još nisu dosegnute!
- Posledica: Vrhunski NLP rezultati mogu se postići samo sa ogromnim hardverskim resursima.

Napomena

- U nastavku je opisan GPT-3 model koji je bio aktuelan 2020. godine.
- Svrha opisa bila je da ilustruje da je dalji pravac razvoja NLP posevćen velikim jezičkim modelima koje može da razvija samo mali broj kompanija u svetu.
- Kao što svi znate od tada su se pojavila dva superiornija modela ChatGPT (GPT-3.5) i GPT-4, što samo potvrđuje da je to budućnost (i sadašnjost) razvoja NLP oblasti.

GPT – 3

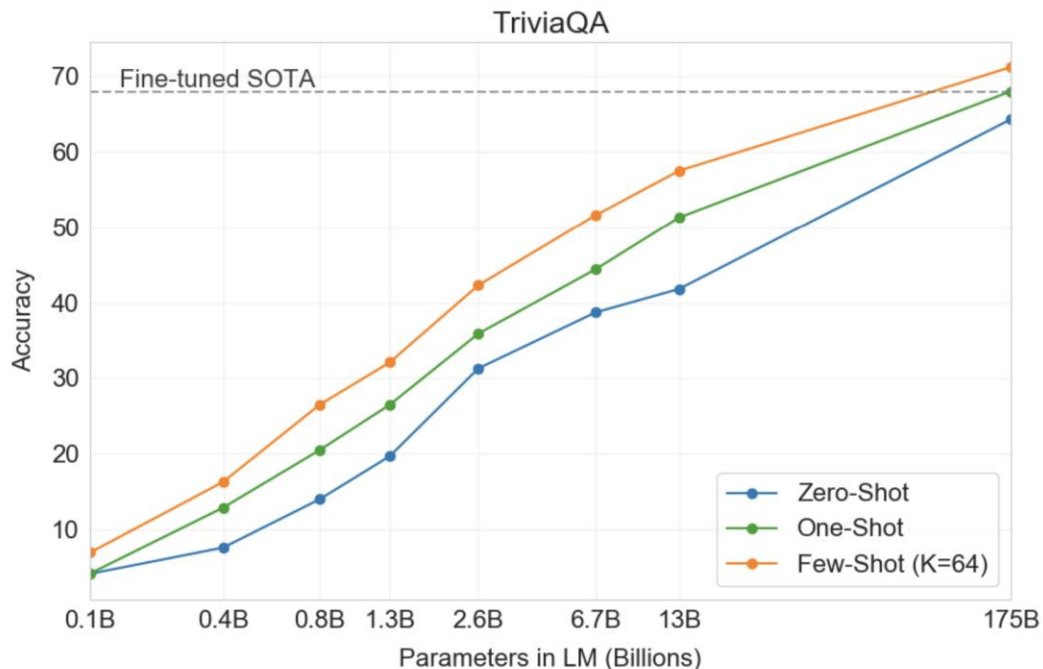
- GPT – 3 je ogroman transformer model koji je razvila kompanija OpenAI.
- Ima ~170 milijardi parametara, obučen je na korpusu od ~500 milijardi tokena.

Microsoft-built AI Supercomputer

- NVIDIA V100 GPUs in a high-bandwidth cluster
 - 285,000 CPU cores
 - 10,000 GPUs
 - 400 gigabits per second network connectivity for each GPU server
 - Trained on cuDNN accelerated PyTorch models
- Would require 355 years and \$4,600,000 train on cheapest GPU cloud

GPT – 3 - rezultati

- Rezultati na Super GLUE većinom nisu bili bolji od najboljih modela.
- Međutim, GTP – 3 zahteva manji obučavajući skup.

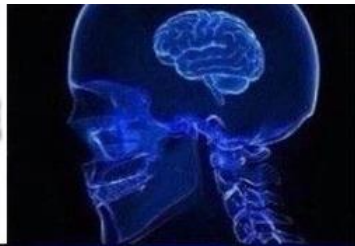


GPT – 3 - rezultati

- Oblast u kojoj se GPT – 3 pokazao kao ubedljivo najbolji model ikada je generisanje teksta.
- Ako se ima u vidu da se skoro svaki problem može formulirati kao pitanje na koji model generiše odgovor GPT – 3 pokazao se kao neverovatno moćan model.
- Pojava GPT – 3 daje nam bljesak potencijalne budućnosti u kojoj da bi rešili problem na računaru ne moramo da napišemo kod, niti da prikupimo dataset, već samo da postavimo pravo pitanje.

<https://twitter.com/karpathy/status/1273788774422441984/photo/1>

**PRE-SOFTWARE:
SPECIAL-PURPOSE
COMPUTER**



**SOFTWARE 1.0:
DESIGN
THE ALGORITHM**



**SOFTWARE 2.0:
DESIGN
THE DATASET**



**SOFTWARE 3.0:
DESIGN
THE PROMPT**



GPT – 3 - <https://gpt3examples.com/>

(Suhail CS) Wear Mask - Save Lives
@ChinyaSuhail

@gdb @smdcmc @awscloud

When GPT-3 Meets DevOps 🤖

** create, deploy, list, and delete any services on AWS
using conversational plain English **

Bootstrapped with @sh_reya's gpt-3 sandbox ❤️
Working on a end-end pipeline with @snpranav

#OpenAI #GPT3 #DevOps #AWS



0:48 13.9K views

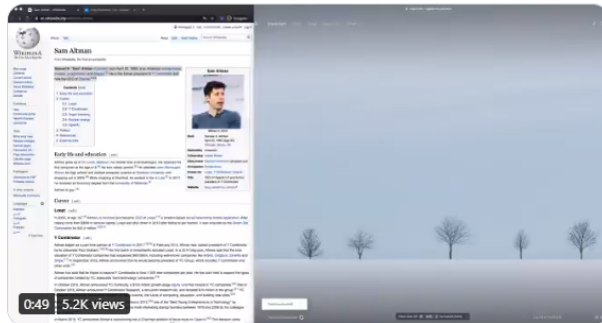
Yigit Ihlamur
@yihlamur

Replying to @yihlamur

The use-cases are endless.

I created an entity extraction demo in less than 10 minutes. Usually, this work requires a significant engineering effort and machine learning expertise.

I can't wait to see what entrepreneurs will build in the next couple of months.



4:20 AM · Jul 26, 2020 · Twitter Web App

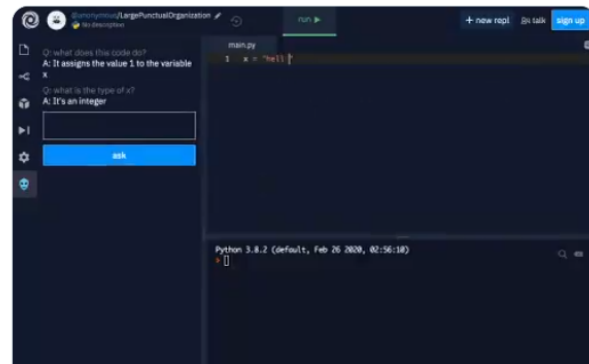
Amjad Masad
@amasad

Reading code is hard! Don't you wish you could just ask the code what it does? To describe its functions, its types.

And maybe... how can it be improved?

Introducing: @Replit code oracle 🧙

It's crazy, just got access to @OpenAI API and I already have a working product!



GPT – 3 - <https://gpt3examples.com/>



Paras Chopra
@paraschopra

I made a fully functioning search engine on top of GPT3.

For any arbitrary query, it returns the exact answer AND the corresponding URL.

Look at the entire video. It's MIND BLOWINGLY good.

cc: @gdb @npew @gwern



Antonio Gómez
@Antonio_GomezM

I managed to develop a GPT-3 demo that produces generative code snippets of the [#threejs](#) JavaScript API by textually describing the elements and their parameters needed to create a [#webgl](#) 3D scene.

GPT-3 and THREE.js API

Can you please create an emp



@Antonio_GomezM

GPT - 3 <https://www.gwern.net/GPT-3>

- GPT - 3 nastavlja da piše tekst članka u kome ga kritikuju.
- Tekst koji nije boldovan generisao je GPT - 3.

“Derp Learning”

By Douglas Hofstadter and Gary Marcus

On recent advances in AI and why they are overblown: computers are just soulless machines, which don't understand anything. They're nothing but pattern recognition and statistical memorization of large corpuses. Artificial intelligence programs like deep learning neural networks may be able to beat humans at playing Go or chess, or doing arithmetic, or writing Navy Seal cypypasta, but they will never be able to truly

... understand the underlying meaning of things. Additionally, we still don't have programs that can truly imagine or create new things all by themselves in a natural way, but rather just brute force through random strings of garbage until they happen to hit on the correct solution to the current problem. We still lack a Rosetta Stone for the universe and don't really know what things like light, sound, matter, gravity, life, free will, and consciousness are.

GPT – 3 - Šta nam je GPT-3 pokazao?

- Granice još uvek nisu dosegnute.
- Čeka se sledeći ogroman model...
- Attention mehanizam može neverovatno dobro da imitira razumevanje jezika.
- Transformer je trenutno najveći korak ka AGI, ali definitivno nije rešenje tog problema.



Geoffrey Hinton
@geoffreyhinton

Extrapolating the spectacular performance of GPT3 into the future suggests that the answer to life, the universe and everything is just 4.398 trillion parameters.

2:26 PM · Jun 10, 2020 · [Twitter Web App](#)

741 Retweets and comments 3.8K Likes



GPT – 3 - Šta nam je GPT-3 potvrdio?

- Hardver je važniji od algoritama....

Rich Sutton's "Bitter Lesson"

- Simple AI leveraging compute power beats clever AI built using human knowledge
- Deep Blue chess machine based on search
- NLP translation based on n-grams
- Scaling of NLP transformer models
- AlphaGo based on search and self-play

- Gwern: OA5, BigGAN, BiT, ViLBERT, AlphaStar, MetaMimic, StyleGAN, GQN, Dactyl, DD-PPO, Progen, AlphaZero, MuZero



<https://www.gwern.net/newsletter/2020/05>

<http://www.incompleteideas.net/Incldeas/BitterLesson.html>

U nastavku predavanja...

1. Na koje sve načine možemo reprezentovati tekst.
2. Šta sve možemo da uradimo pomoću NLP tehnika i kako to možemo da uradimo:
 1. Pomoću gotovih modela (alata)
 2. Štelovanjem postojećih modela
 3. Od nule

Načini za reprezentovanje teksta

Od one-hot vektora do GPT-3

One-hot vektori

- U tradicionalnom NLP svakoj reči odgovarao je jedan binaran vektor:

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

- Dimenzionalnost vektora odgovara veličini rečnika.
- One-hot - 1 na poziciji (koordinati) koja odgovara datoj reči, a 0 na svim ostalim pozicijama.

One-hot vektori – Šta je problem?

- Varijabilnost jezika predstavlja najveći problem.
- Na primer, vaš upit je “hoteli u Novom Sadu“, a verovatno očekujete da će u rezultatima da se nađu i moteli, ali vektori
$$\text{motel} = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0]$$
$$\text{hotel} = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$
- su ortogonalni....Kod one-hot vektora ne reči ne postoji prirodan način za računanje sličnosti.
- Rešenje: umesto da tražimo bolji metod za poređenje one-hot vektora, tražimo bolji metod za formiranje vektora.

Distribuciona semantika

- Značenje reči određujemo pomoću reči koje se često nalaze oko nje.



“You shall know a word by the company it keeps” (J. R. Firth 1957: 11)

*...government debt problems turning into **banking** crises as happened in 2009...*

*...saying that Europe needs unified **banking** regulation to replace the hodgepodge...*

*...India has just given its **banking** system a shot in the arm...*

- Reč reprezentujemo kao vektor koji zavisi od vektora reči koje se često nalaze oko nje.

Distribuciona semantika – vektori reči

- Vektor za neku reč formiramo tako da su slični vektorima reči koje se često nalaze u okolini (kontekstu) te reči.

banking =

$$\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

- Ovakve reprezentacije se često nazivaju *word embeddings*.

Distribuciona semantika – vektori reči

- Primer 2d projekcije vektora dobijenih distribucionom semantikom

banking =

$$\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

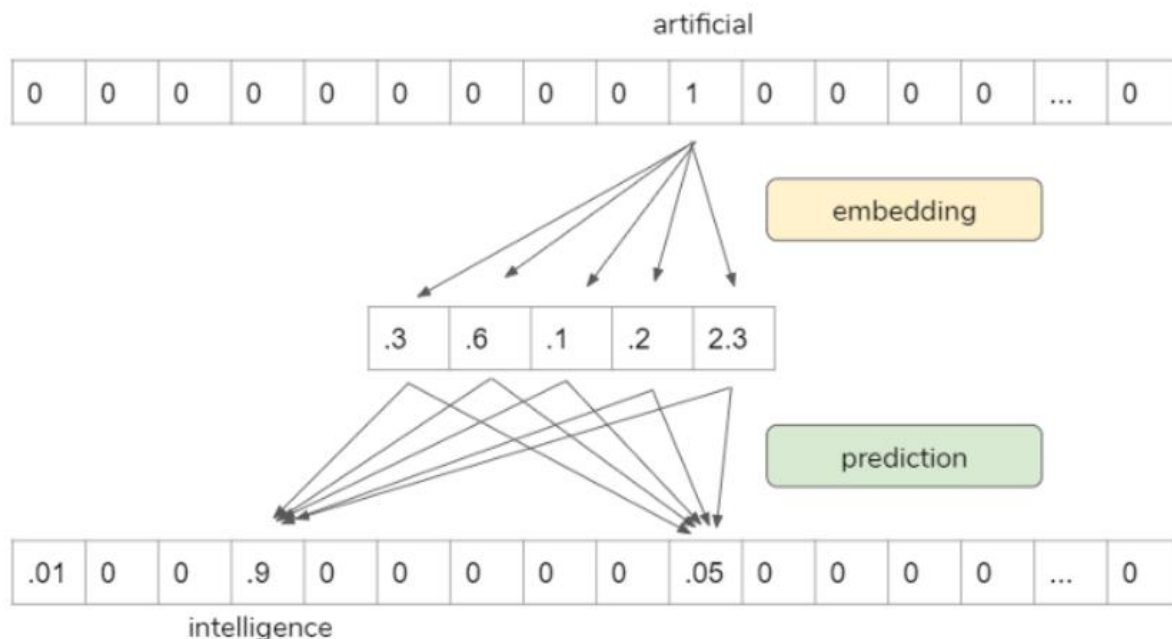

<https://projector.tensorflow.org/>

Word2vec

- Jedan od najpoznatijih modela distribucione semantike.
- Osmišljen je 2013 (Mikolov et al. 2013) ali se koristi i danas.
- Ideja:
 - Prolazimo kroz korpus sa klizećim prozrom za reči (okolina reči)
 - Za svaku reč rešavamo jedan od sledeća dva predikciona problema:
 - Predvideti tu reč na osnovu reči iz okoline (**C**ontinuous **B**ag **O**f **W**ords varijanta)
 - Predvideti okolne reči na osnovu te reči (Skip-gram varijanta)
 - Koristimo obučeni model da proizvedemo vektore reči.

Word2vec

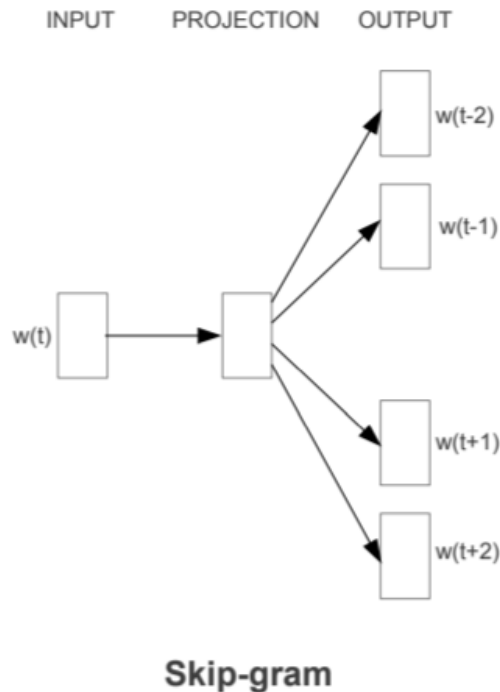
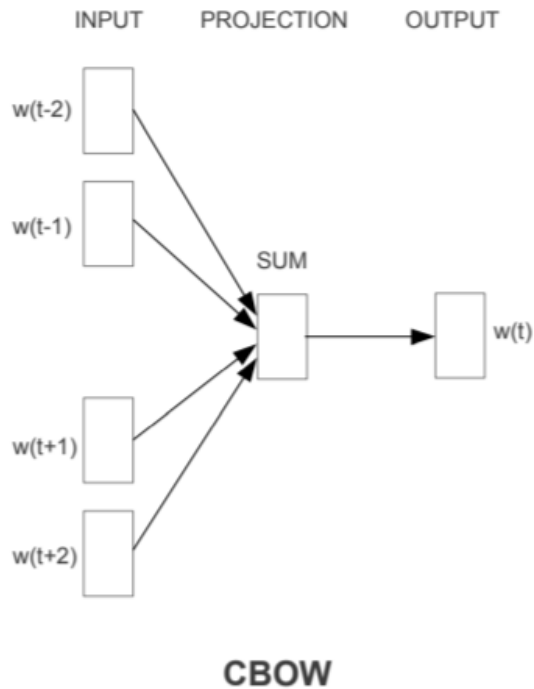
- Tipično se implementira se kao neuronska mreža sa jednim skrivenim slojem.



- Vrednosti skrivenog sloja su vektori pomoću kojih reprezentujemo reči (*word embeddings*).

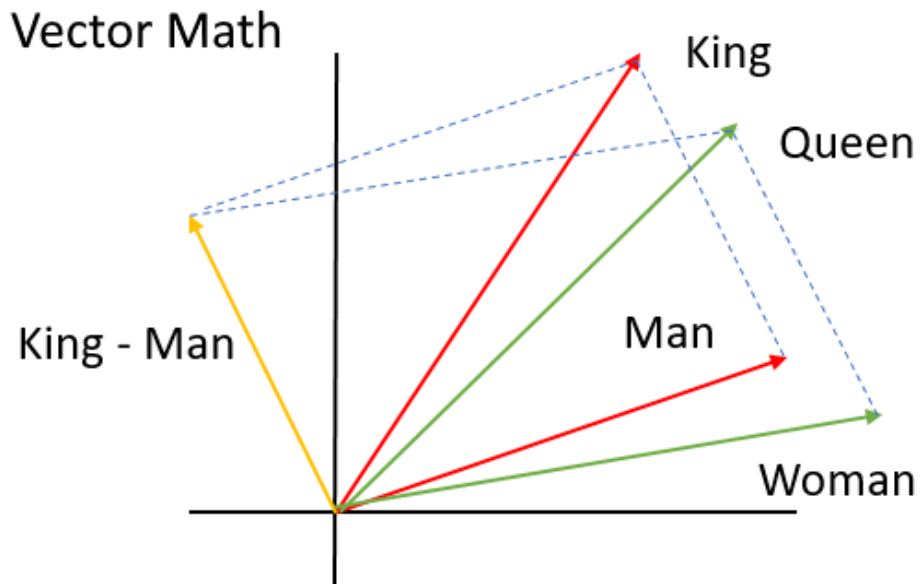
Word2vec

- Skip-gram vs. CBOW



Word2vec

- Pored poređenja sličnosti, možemo raditi i vektorsku algebru u prostoru vektora reči. Čuveni primer: $\text{King} - \text{Man} + \text{Woman} = \text{Queen}$



Glove

- Jedan od najpoznatijih metoda pored word2vec, kreiran na Stenfordu (Pennington et al. 2014]).
- Oslanja se na matricu brojeva međusobnog zajedničkog pojavljivanja reči (*co-occurence matrix*).
- Ta matrica se formira za okolinu (prozor oko) reči.
- Na sledećem slajdu dat je primer jedne matrice za korpus od tri rečenice.

Glove

- Korpus: “I like deep learning. “, “I like NLP.“, “I enjoy flying.“

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

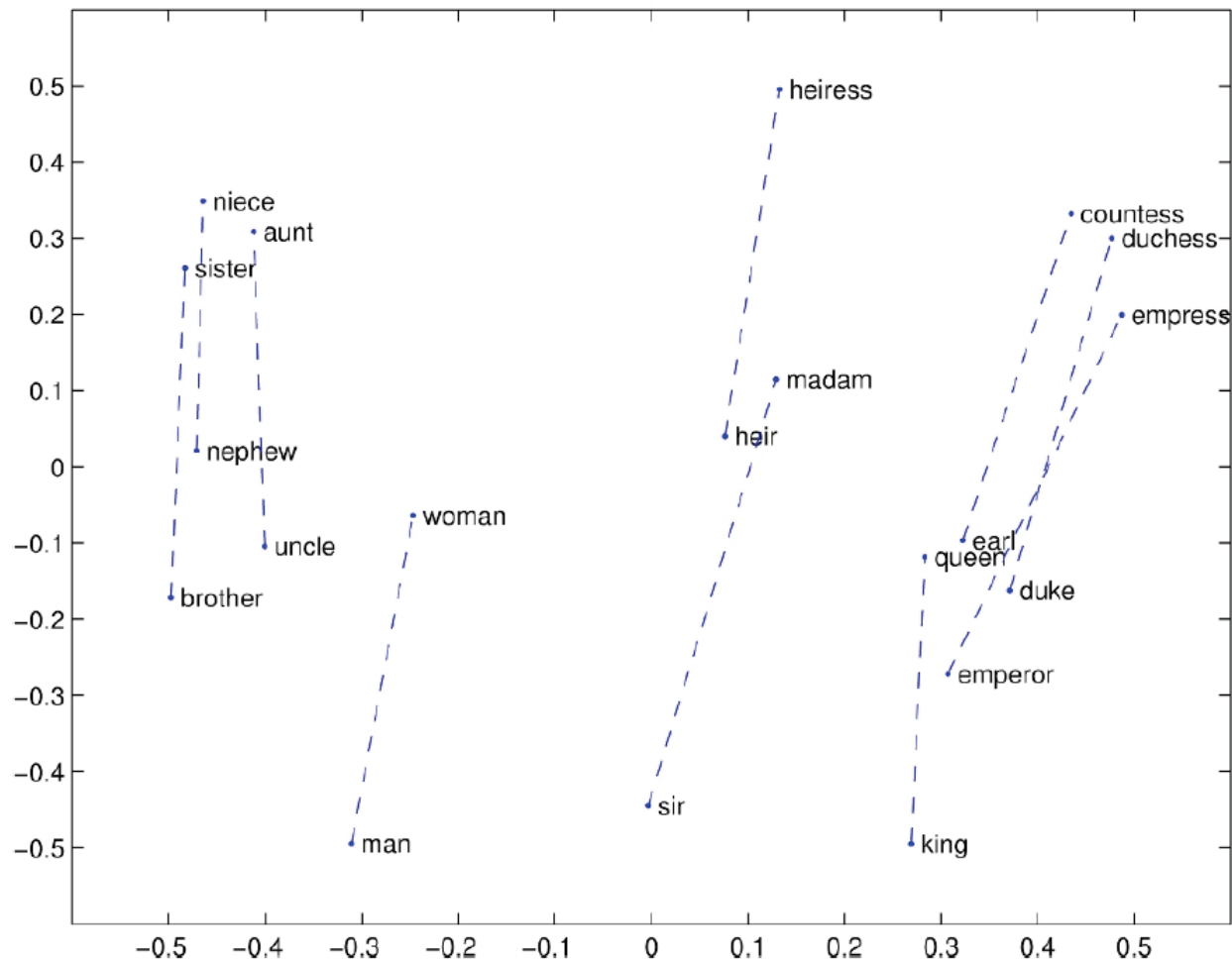
Glove

- Ideja na osnovu koje se formiraju Glove vektori je:
- rastavljanje co-occurrence matrice na delove pomoću metode koja se zove *Singular Value Decomposition* (SVD).
- Iako je ideja primene SVD dobra, sam postupak SVD ispostavio se kao problematičan u praksi.
- Autori Glove pristupa su zato SVD sveli na optimizacioni problem.

Glove

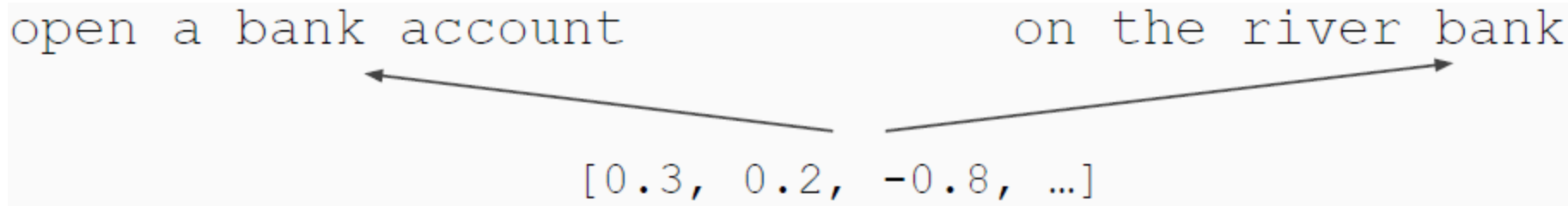
- Autori Glove pristupa su zato SVD sveli na optimizacioni problem.
- Time su napravili jako robustan i skalabilan metod (u odnosu na veličinu korpusa).
- Primeri Glove vektora dati su na sledećem slajdu.

Glove



Word2vec i Glove - mane

- Najveća mana ovih metoda je što za svaku reč postoji samo jedan vektor na koji se mapira.
- Kao što znamo, jedna reč može imati više značenja, koja zavise od konteksta u kojima se reč nalazi.



Word2vec i Glove - rezime

- Pored ove dve metode postoje i mnoge druge slične, kao npr. FastText (Bojanowski et al. 2017) i Flair (Akbik et al. 2018).
- Flair je takođe i okruženje koje se stalno unapređuje i koje nudi razne vektorske preprezentacije, ali njihove kombinacije (<https://github.com/flairNLP/flair>).
- Word2vec (i ostali) još uvek imaju značajno mesto u otkrivanju semantike reči jer se koriste kao ulaz za BERT i slične metode o kojima govorimo u nastavku.

Kontekstualne reprezentacije

- Jedna reč može imati više različitih reprezentacija (vektora).
- Vrednosti vektora zavise od konteksta u kome se reč nalazi.

[0.9, -0.2, 1.6, ...]



open a bank account

[-1.9, -0.4, 0.1, ...]

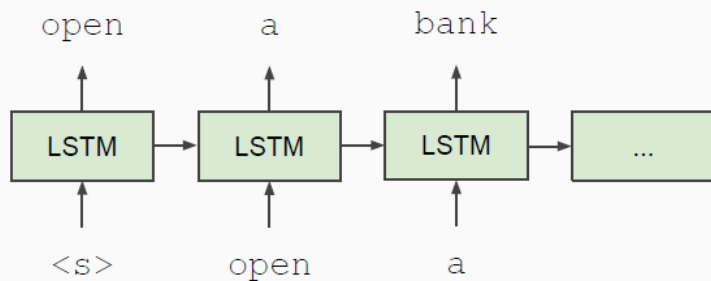


on the river bank

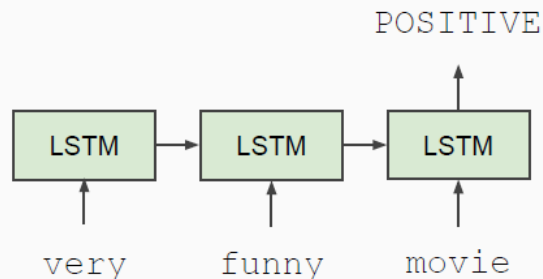
Kontekstualne reprezentacije

- Prve metode se pojavljuju 2015 u Google.
- Koristi se LSTM mreža koja se obučava da predvidi sledeću reč.

Train LSTM Language Model



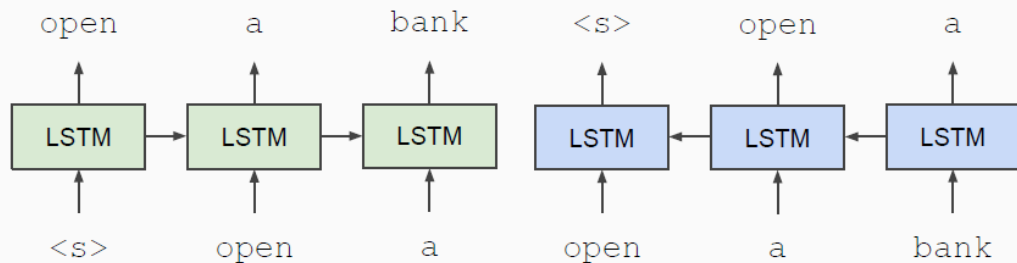
Fine-tune on Classification Task



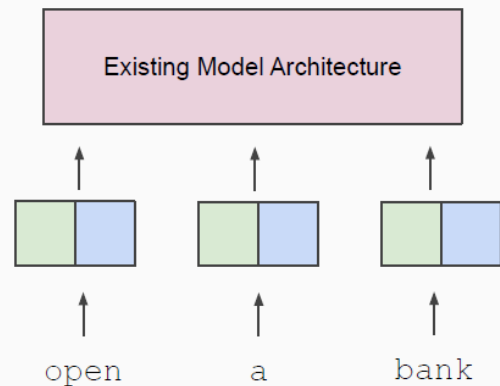
Kontekstualne reprezentacije

- Nakon toga 2017. pojavljuje se ELMo model (Peters et al.).
- Takođe LSTM mreža, ali se model obučava i sa obrnutim redosledom reči.

Train Separate Left-to-Right and Right-to-Left LMs

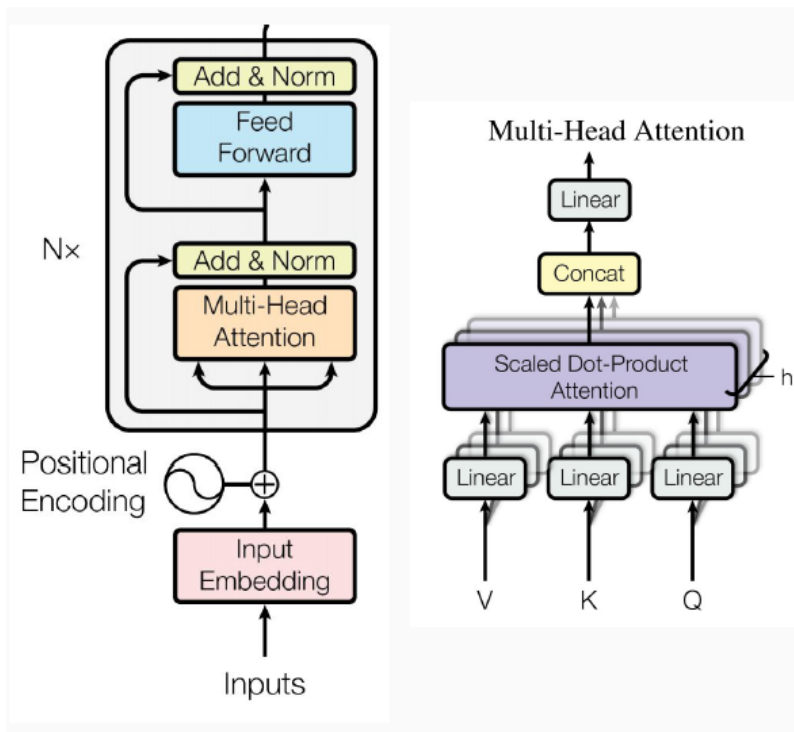


Apply as “Pre-trained Embeddings”



Kontekstualne reprezentacije

- Korak koji je uneo revoluciju je pojava transformer modela koji koristi mehanizam samo-pažnje (*self-attention*). (Vaswani 2017 – Google Brain).



BERT

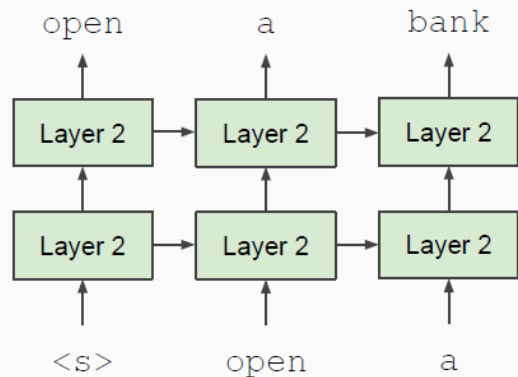
- BERT model je nastao u Google kao primena transformer arhitekture i novog načina obučavanja (Devlin et al. 2018)
- Do BERT-a modeli su obučavani samo u jednom pravcu (sa leva na desno ili obrnuto).
- Bilo je pristupa koji predikciju sledeće reči vrše i sa leva na desno, i sa desna na levo, ali su tada obučavana dva zasebna uni-direkciona model, a ne jedan bi-direkcioni model.

BERT

- Zašto su uni-direkcionni modeli bili dominantni do BERT modela?
- Zato što je obučavanje uvek bilo takvo da se predviđa sledeća reč.
- Ako je model bi-direkcionni ne može se primeniti takav način obučavanja jer se reči “vide međusobno”.

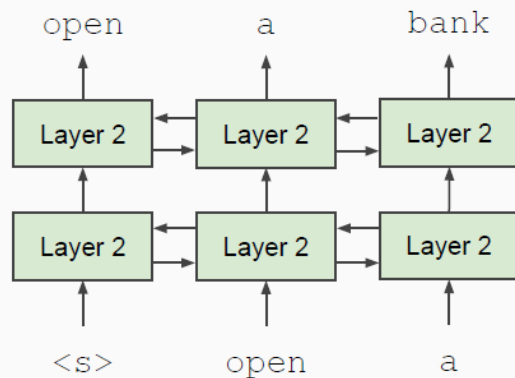
Unidirectional context

Build representation incrementally



Bidirectional context

Words can “see themselves”



BERT

- Problem: reči “se vide međusobno”.
- Rešenje: promeniti način obučavanja.
- BERT: sakriti (maskirati) $k\%$ reči u rečnici i tražiti od modela da ih predvidi.
- Konkretno BERT koristi $k=15$.

store gallon

↑ ↑

the man went to the [MASK] to buy a [MASK] of milk

BERT

- Kao dodatno poboljšanje kvaliteta modela BERT se obučavao na još jedan način.
- Predviđanje sledeće rečenice na osnovu prethodne:

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

BERT

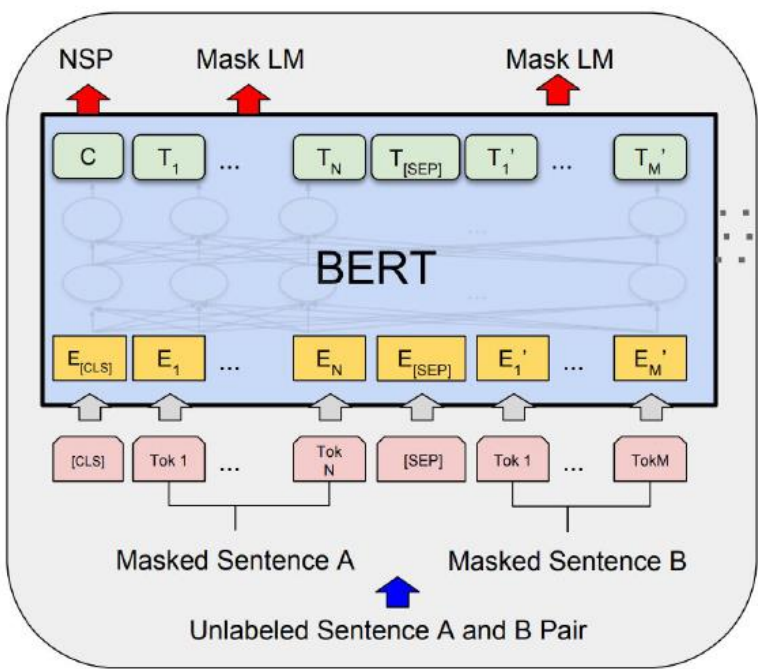
- BERT je obučavan na korpusima: Wikipedia (2.5B reči) i BookCorpus (800M)
- Obučavanje: 1M koraka (~40 epoha) 4 dana na 4x4 ili 8x8 TPU procesorima
- Dva modela:
 - “manji”: BERT-Base: 12-layer, 768-hidden, 12-head
 - veći: BERT-Large: 24-layer, 1024-hidden, 16-head
- Poslednji skriveni sloj (*hidden*) koristi se za reprezentovanje reči (dimenzionalnosti 768 za Base i 1024 za Large).

BERT

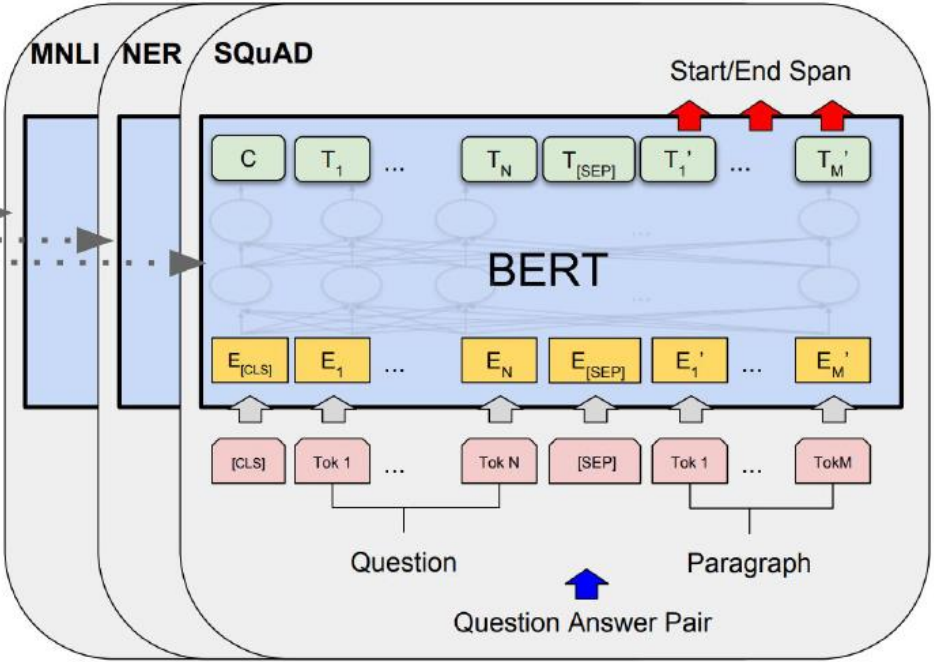
- Nakon obučavanja, BERT je dodatno obučen, tj. fino naštelovan (*fine-tuning*) na različitim NLP zadacima.
- Za prvo obučavanje (*pre-training*) BERT-a (i svih ostalih modela za reprezentovanje reči) korišćen korpus bez ikakvih oznaka klasa, odnosno korišćeni su ne-anotirani podaci.
- Fino-štelovanje podrazumeva upotrebu reprezentacija koje je BERT naučio u pre-training fazi da bi ga sada naučili da obavlja neki specifičan zadatak (npr. odgovaranje na pitanja).

BERT

- *Pre-training* | *fine-tuning*



Pre-training



Fine-Tuning

BERT

- BERT je postigao bolje rezultate od state-of-the-art modela na svim GLUE (preteča SuperGLUE) zadacima.
- Od tada je nastalo mnogo varijacija BERT modela (neke su pomenute na početku predavanja).
- Međutim, varijacije su male i svode se na promenu korpusa, trajanja obučavanja i veličine modela.
- Od BERT modela nije kreirana arhitektura koja bi unela novu revoluciju u NLP.

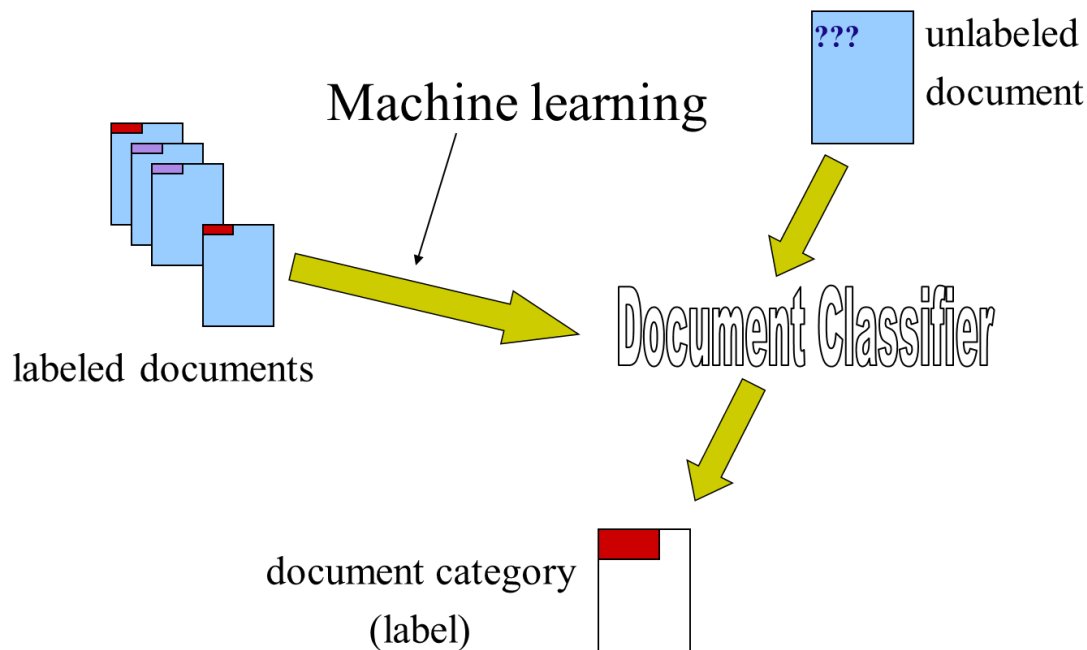
U nastavku predavanja...

1. Na koje sve načine možemo reprezentovati tekst.
2. **Šta sve možemo da uradimo pomoću NLP tehnika i kako to možemo da uradimo:**
 1. Pomoću gotovih modela (alata)
 2. Štelovanjem postojećih modela
 3. Od nule

Zadaci koje možemo obaviti pomoću NLP (*NLP taks*)

Klasifikacija teksta

- Jedan od prvih i lakših zadataka.
- Može se raditi različitim nivoima (dokument, pasus, rečenica itd.)
- Nivo dokumenta je tipičan.



Klasifikacija teksta

- Dokumenti se tipično reprezentuju kao vektori.
- Najčešći način je *Bag-Of-Words* (BOW) metod.
- Formiramo rečnik pomoću reči iz korpusa i opciono ga filtriramo:
 - uklanjamo stop-reči (reči koje se često pojavljuju, a nemaju semantiku)
 - reči svodimo na svoj korenski oblik (stemming i lemmatization)
 - izbacujemo ne-frekvente reči
 - dodajemo fraze od 2 (bi-grami), 3 (uni-grami) itd. reči
 - sadržavano samo reči koje imaju neku statističku vezu sa klasama (npr. pomoću chi-squared mere).
 - ...

Klasifikacija teksta

- Kada imamo rečnik, formiramo vektor za svaki dokument.
- Komponente vektora su reči i ima ih koliko ima reči u rečniku.
- Vrednosti komponenti određujemo na različite načine:
 - binaran vektor (1 ako je reč u dokumentu, inače 0)
 - vektor frekvencija (c – gde je c broj pojavljivanja reči u dokumentu)
 - tfidf – najčešće korišćena mera (detaljno na sledećem slajdu)

Klasifikacija teksta

- Tfidf mera da je veće vrednosti rečima koje su specifične baš za dati tokument (da ponovimo svaki dokument je poseban vektor).

$$tfidf(w) = tf(w) \cdot \log\left(\frac{N}{df(w)}\right)$$

Reč je značajna za dokument ako se često javlja u njemu

Reč je značajna ako je specifična za ovaj dokument, tj. ne javlja se puno u drugim dokumentima

- *tf* – *term frequency*, koliko se puta reč javlja u dokumentu
- *df* – *document frequency*, broj dokumenata koji sadrže tu reči u korpusu.
- *N* – broj dokumenata u korpusu.

Klasifikacija teksta

- Kada imamo vektor problem je sveden na klasičan klasifikacioni problem iz mašinskog učenja.
- Savremene tehnike za reprezentaciju reči kao BERT mogu se koristiti i u ovom slučaju.
 - Na primer, možemo da uzmemo prosek vektora za svaku reč kao reprezentaciju dokumenata.
 - Postoje i gotovi doc2vec pristupi.
- Sve poznate biblioteke za NLP nude načine za formiranje vektora dokumenata. Dve poznate:
 - nltk (<https://www.nltk.org/>)
 - Spacy <https://spacy.io/>

Klasifikacija teksta

- Ako znamo kako da dobijemo vektore i da obučimo klasifikacione modele, šta nam još nedostaje da bi rešili neki naš klasifikacioni problem?
- Nedostaje nam korpus za obučavanje.
- Prisjetite se prethodnog predavanja i priče o prikupljanju i anotiranju obučavajućeg skupa.
- Sve to važi i ovde.

Određivanje sentimenta teksta

- Zadatak je odrediti da li je u tekstu izražen **pozitivan** ili **negativan** sentiment (ponekad i **neutralan**).
- Može se raditi na nivou dokumenta, pasusa, rečenice.
- Često se primenjuje na objavama sa društvenih mreža (najviše na Twitter-u).



Određivanje sentimenta teksta - metodologije

- **Rečnici:**
- Formirati (ili pronaći) rečnike za pozitivne, negativne i neutralne reči.
- Prebrojati koliko se reči iz dokumenta nalaze u rečnicima i doneti odluku o sentimentu.
- Jednostavan, ali i najlošiji pristup.
- Bolje je da se koristi u kombinaciji sa drugim pristupima.

Određivanje sentimenta teksta - metodologije

- **Pravila:**
- Formirati (ili pronaći) ručno napisana pravila na osnovu kojih se određuje sentiment.
- Na primer if “This is great” in text: sentiment=positive.
- Mukotrpan posao, koji može da radi dobro za uže domene.
- Može se koristiti u kombinaciji sa drugim pristupima.

Određivanje sentimenta teksta - metodologije

- **Mašinsko učenje:**
- Svodi se na problem na klasifikacije teksta.
- Koristi se neka od vektorskih reprezentacija.
- Vektor se može proširiti rezultatima primene rečnika i pravila.
- Kod rada sa tekstovima sa društvenih mreže treba obratiti pažnju na emotikone i korekciju spelovanja (*spelling correction*).

Određivanje sentimenta teksta - napomene

- Na nižim nivoima (npr. rečenica) zadatak nije baš tako lak:
 - “The acting is superb, the CGI are amazing, but overall the movie sucks.”
 - “Your perfume is great, I suggest you use it in your apartment with the windows closed.”
- Treba uzeti u obzir negaciju, sarkazam itd.
- Međutim, ako vam treba neka agregacija sentimenta (npr. šta ljudi na tviteru misle o nekoj temi) statistika će „ispeglati“ evetualne greške
 - poneki sarkastičan tvit koji ste promašili neće promenti opšti sentiment.

Određivanje sentimenta teksta - napomene

- Postoji mnogo gotovih alata za engleski jezik:
 - TextBlob <https://textblob.readthedocs.io/en/dev/>
 - StanfordCoreNLP <https://stanfordnlp.github.io/CoreNLP/>
 - Vader <https://github.com/cjhutto/vaderSentiment>

Određivanje sentimenta teksta - aspekti

- Aspekt je deo teksta (tipično rečenice) na koji se sentiment odnosi.
- Nekada je aspekt agregiran na viši nivo i poznat unapred.
 - Na primer, ako obrađujete recenzije filmova ne zanima vas šta se o kom glumcu misli na nivou rečenice već su predmet obrade filmovi.
- Kada nam je potrebne finija analiza, npr. “Kamera je odlična, ali je baterija loša“, moramo da odredimo aspekte i povežemo ih sa sentimentom.
- Analiza sentimente zasnovnana na aspektima je težak zadatak.

Određivanje sentimenta teksta - aspekti

- Postoje različite metodologije za detekciju aspekata.
- Načešće se svode na:
 - Ekstrakciju ključnih reči ili fraza iz teksta
 - Prepoznavanje imenovanih entiteta u tekstu
- U nastavku obrađujemo obe metodolgije.

Ekstrakcija ključnih fraza iz teksta

- Zadatak je da se u tekstu otkriju reči ili fraze koje najbolje reprezentuju temu o kojoj se govori u tekstu.

Natural language processing (NLP) is a subfield of linguistics , computer science , and artificial intelligence concerned with the interactions between computers and human language , in particular how to program computers to process and analyze large amounts of natural language data . Challenges in natural language processing frequently involve speech recognition , natural language understanding , and natural-language generation.

<http://www.nactem.ac.uk/software/terminology>

- Pristupi se zasnivaju na kombinaciji leksičke, sintakse i statistike.
 - Na primer ključne reči su tipično imenice pa treba da znamo vrste reči.
 - Tipično su frekventne u datom tekstu, ali ne toliko u ostalim tekstovima itd.
- Pomoću svih do sada navednih alata direktno ili indirektno mogu se dobiti ključne fraze
 - na primer za TexBlob <https://textblob.readthedocs.io/en/dev/quickstart.html>

Ekstrakcija imenovanih entiteta

- Zadatak koji se tipično naziva NER (*Named Entity Recognition*).
- Cilj je pronaći entitete od interesa u tekstu.

Alan Mathison **PERSON** Turing OBE FRS (/ˈtʃʊərɪŋ/; 23 June 1912 **DATE** – 7 June 1954 **DATE**) was an English mathematician, computer scientist, **logician** **NORP**, **cryptanalyst** **PERSON**, philosopher, and theoretical biologist.[6][7] Turing was highly influential in the development of theoretical computer science, providing a formalisation of the concepts of **algorithm** **ORG** and computation with the **Turing** **PERSON** machine, which can be considered a model of a general-purpose computer.[8][9][10] **Turing** **GPE** is widely considered to be the father of theoretical computer science and artificial intelligence.[11] Despite these accomplishments, he was never fully recognised in his home country during his lifetime due to the prevalence of **homophobia** **GPE** at the time and because much of his work was covered by the Official Secrets Act.

During the Second World War, **Turing** **PERSON** worked for the Government Code and Cypher School (GC&CS) at **Bletchley Park** **GPE**, **Britain** **GPE**'s codebreaking centre that produced **Ultra** **ORG** intelligence. For a time he led **Hut 8** **PERSON**, the section that was responsible for **German** **NORP** naval cryptanalysis. Here, he devised a number of techniques for speeding the breaking of **German** **NORP** ciphers, including improvements to the pre-war **Polish** **NORP** bombe method, an electromechanical machine that could find settings for the **Enigma** **PRODUCT** machine.

Ekstrakcija imenovanih entiteta

- Ovo je zadatak na kome se puno radi u svim domenima.
- Trenutno je bio-medicinski domen jako aktuelan
 - prepoznavanje gena, proteina itd. u načunim radovima.
 - postoje organizacije koje se samo time bave uz ručnu proveru (*curation*).
- Naravno procesiranje vesti i objava na društvenim mrežama je uvek aktuelno i od interesa kako kompanijama tako i državama.
 - Rezultati se npr. mogu povezati sa sentimentom.

Ekstrakcija imenovanih entiteta

- Sve metodologije koje smo pomenuli za određivanje senitimenta mogu se koristiti i ovde:
 - rečnici, pravila i mašinsko učenje.
- Nekada je Conditional Random Fields (CRF) metod bio dominantan.
- Sada dominiraju metode dubokog učenja
 - modeli kao što su BERT mogu se lako fino-doštlovati da prepoznaju entitete

Ekstrakcija imenovanih entiteta

- Postoje naravno gotovi alati:
 - Spacy se jako puno koristi
 - StanfordCoreNLP je takođe popularan...
- Ako želite da napravite model da prepozna neke specifične entitete trenutno vam je najteži zadatak da napravite korpus.
- Anotiranje entiteta nije toliko jednostavno kao anotiranje dokumentata. Postoje alati za anotiranje (npr. <http://keighrim.github.io/mae-annotation/>)
 - Treba imati u vidu da se više ljudi ne slaže uvek oko toga šta je entitet
 - Objektivnost korpusa tipično se garantuje kroz *Inter Annotator Agreement*

Ekstrakcija tema iz teksta

- Pretpostavljamo da je dokument mešavina nekih tema i automatski pokušavamo da odredimo te teme.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

LDA (*Latent Dirichlet Allocation*)

- Dokument se generiše tako što bacimo kockicu i izvučemo temu, pa bacimo kockicu i izvučemo reč iz te teme.
- LDA šteliuje generativni model tako da su baš dokumenti koje imamo u korpusu najverovatniji da budu dobijeni u bacanju kockice.
- Rezultat: Teme korpusa, Distribucija tema po dokumentu. Distribucija reči po temi.

Word	Probability
red	0.202
blue	0.099
green	0.096
yellow	0.073
white	0.048
color	0.030
bright	0.029
colors	0.027
brown	0.027
....

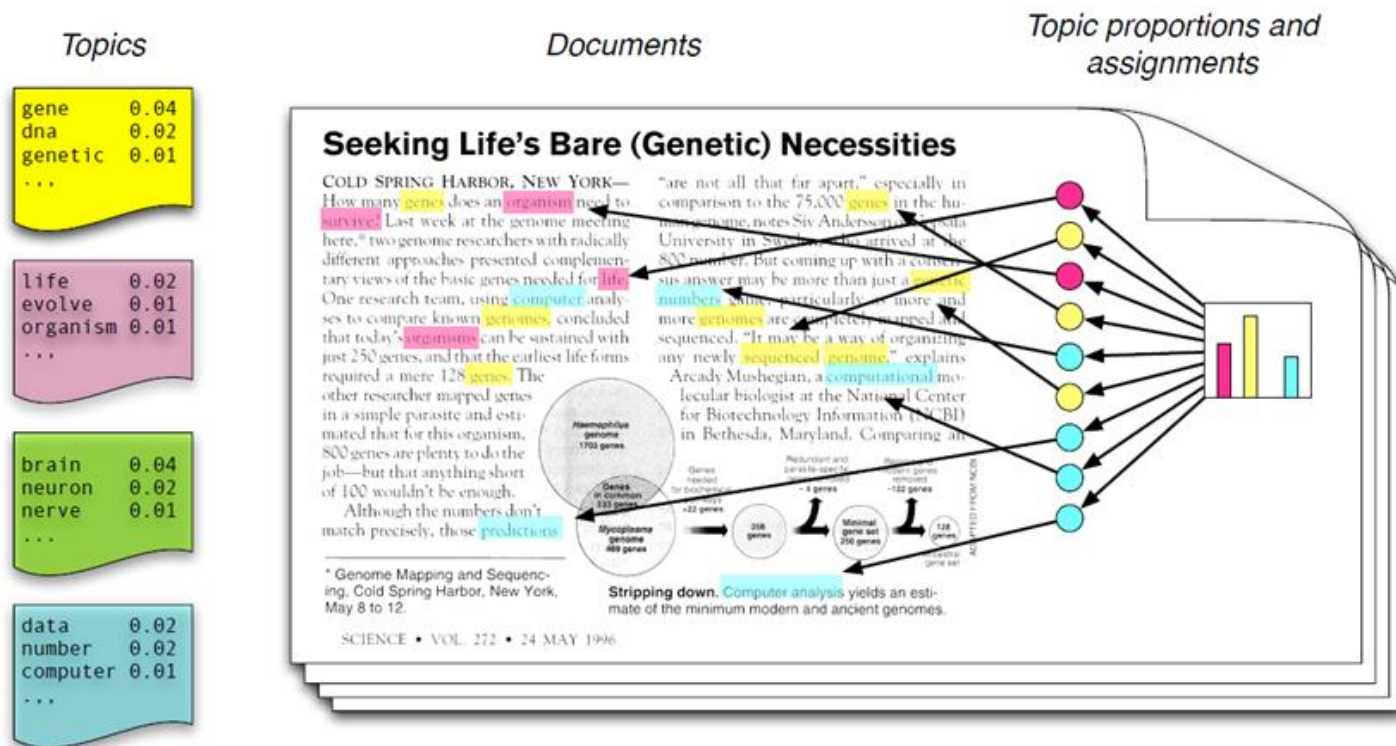
Word	Probability
president	0.129
roosevelt	0.032
congress	0.030
johnson	0.026
office	0.021
wilson	0.021
nixon	0.020
reagan	0.018
kennedy	0.018
....

$P(\text{Words}) \text{ for Doc 1} = 0.4 * \text{Topic 1} + 0.4 * \text{Topic 2} + 0.2 * \text{Topic 3}$

$P(\text{Words}) \text{ for Doc 2} = 0.0 * \text{Topic 1} + 0.8 * \text{Topic 2} + 0.2 * \text{Topic 3}$

LDA

- Generativni model: dokument se generiše tako što semplujemo teme, pa iz sepmovane teme semplujemo reč.



LDA

- LDA metod kao ulaz zahteva samo korpus tekstova.
 - Ne trebaju nam oznake klasa, entiteta ili bilo šta slično.
- Međutim, u praksi je bolje prvo pred-procesirati tekst, konkretno:
 - izbaciti stop reči
 - uraditi stemming ili lematizaciju
- Što bolje uradimo korak pred-proceisranja to će rezultati LDA biti bolji.
- Suština je u tome da pred-proceisranjem izbacimo što više reči koje nemaju semantiku koja nam je od interesa.

LDA

- Polularni alati za LDA:
 - Gensim <https://radimrehurek.com/gensim/>
 - Stanford Topic Modeling Toolbox
<https://nlp.stanford.edu/software/tmt/tmt-0.4/>
 - MALLET <http://mallet.cs.umass.edu/>

Primeri

iz oblasti NLP

SIAP projekata prethodnih generacija

„Automatska detekcija stavki menija unutar teksta recenzija restorana“

- Autori: Igor Trpovski, Marija Joksimović
- Cilj: automatska detekcija pominjanja hrane unutar recenzija restorana i određivanje na koju stavku menija se ono odnosi
- Skup podataka: recenzije restorana sa www.donesi.com
 - ručno anotiran od strane autora projekta
- Modeli: CRF, LSTM, GRU
 - ulazi u RNN mreže: one-hot reprezentacija reči, one-hot reprezentacija reči i karaktera i FastText vektorska reprezentacija reči.

Prikupljanje i filtriranje korpusa

- Prikupljanje pomoću **Selenium**, a parsiranje HTML stranica izvršeno je pomoću biblioteke **BeautifulSoup4**
- Prikupljeno: 169,486 recenzija za 1,658 različitih restorana.
- Brzo ručno filtriranje da bi zadržali samo one recenzije restorana koje u svom naslovu ili tekstu sadrže jedno ili više pominjanja hrane. Rezultat: 20,079 filtriranih recenzija.
- Pored tog još dva koraka pred-procesiranja
 - Konverzija ćirilice u latincu.
 - Uklanjanje dijakritika iz latinice.

Anotiranje korpusa

- Upotrebljen je alat **MAE**. Kreirana je anotaciona shema. Anotirano je 10, 000 recenzija

<http://keighrim.github.io/mae-annotation/>

```
<!ENTITY name "ReviewsTask">

<!ELEMENT B-FOOD ( #PCDATA ) >
<!ATTLIST B-FOOD spans #IMPLIED >

<!ELEMENT I-FOOD ( #PCDATA ) >
<!ATTLIST I-FOOD spans #IMPLIED >

<!ELEMENT L-FOOD ( #PCDATA ) >
<!ATTLIST L-FOOD spans #IMPLIED >

<!ELEMENT U-FOOD ( #PCDATA ) >
<!ATTLIST U-FOOD spans #IMPLIED >
```

```
<?xml version="1.0" encoding="UTF-8" ?>

<ReviewsTask>
<TEXT><![CDATA[kalcona i palacinka
Hrana je ok, brzo dostavljaju.]]></TEXT>
<TAGS>
<U-FOOD id="U0" spans="0~7" text="kalcona" />
<U-FOOD id="U1" spans="10~19" text="palacinka" />
</TAGS>
</ReviewsTask>
```

CRF

- ML model za tagovanje sekvenci. **CRF++** implementacija.
- Svaki token je poseban vektor osobina.

CRF Atributi Tokena	Tokeni			
	Piletina	sa	indijskim	orahom
POS tag	Ncfsn	Si	Agpmsiy	Ncmsi
Lemma	piletina	sa	indijski	orah
Stem	pileti	sa	indijsk	orah
Mala slova	piletina	sa	indijskim	orahom
Valika slova	PILETINA	SA	INDIJSKIM	ORAHOM
Valiko početno, ostala mala slova	Piletina	Sa	Indijskim	Oraham
Da li je broj	False	False	False	False
Da li je znak interpunkcije	False	False	False	False
Da li su sva slova velika	False	False	False	False
Da li je početno slovo veliko, a ostala mala	True	False	False	False
Da li je kraj rečenice	False	False	False	True

RNNs

- Modeli:
 - Bidirekcionni LSTM/GRU modeli sa one-hot reprezentacijom reči
 - Bidirekcionni LSTM/GRU modeli sa one-hot reprezentacijom reči i karaktera
 - Bidirekcionni LSTM/GRU sa FastText vektorskim reprezentacijama reči
 - Bidirekcionni LSTM/GRU modeli dodatnim sa CRF slojem
- Alat: **Keras**.

Rezultati

Model	Preciznost	Odziv	F_1 -mera
CRF	0.9359	0.8974	0.9162
Bidirekcioni LSTM sa one-hot kodiranjem reči	0.9090	0.9024	0.9057
Bidirekcioni LSTM sa FastText vektorskim reprezentacijama reči (5 epoha)	0.9294	0.9212	0.9253
Bidirekcioni LSTM-CRF sa FastText vektorskim reprezentacijama reči (10 epoha)	0.9237	0.9307	0.9272

Analiza grešaka

- Neuspelo prepoznavanje pominjanja hrane koje sadrže greške u pisanju.
 - Ovaj tip grešaka najviše se vezuje za CRF model. Na primer, “Capricciossi”, “dressinga”, “butkica”, “KALZONE” i “kachkavalj”...
- Neuspelo prepoznavanje dela pominjanja hrane čiji je tekst sadržan iz više tokena (CRF)

Fraza	Predikcija	Očekivano
ovciji	O	B-FOOD
sir	U-FOOD	L-FOOD

- Netačno prepoznavanje tokena kao delova teksta pominjanja hrane.

Fraza	Predikcija	Očekivano
odusevila	B-FOOD	O
corba	L-FOOD	U-FOOD

Rezime

- Metodološki jedan od najboljih projekata na predmetu.
- Svaki korak je odabran sa odgovarajućom motivacijom, primenjen i kasnije detaljno objašnjen.
- Analiza grešaka je takođe na jako visokom nivou
 - Analiza grešaka je važan korak kod NLP projekata jer se iz tog dela vidi koliko je problem težak i na čemu treba dalje da se radi.
- Projekat je proširen sa detekcijom sentimenta i rezultovao je sa dva master rada.

“Automatsko određivanje tema knjiga pomoću tehnika za procesiranje prirodnog jezika”

- Autor: Vlada Đurđević
- Cilj: odrediti teme knjiga u okviru šireg projekta predikcije popularnosti knjiga
- Skup podataka:
 - obučavajući: *CMU Book Summary dataset* - meta-podaci o 16,559 različitih knjiga
 - test: podaci o knjigama prikupljeni sa sajta Goodreads, za 200 najpopularnijih knjiga iz svake od prethodnih 100 godina
- Modeli: različite implementacije LDA modela

Pred-procesiranje korpusa 1/2

- Tokenizacija
- Uklanjanje stop-reči
 - lista stop-reči iz **NLTK** platforme, dodatno proširena rečima koje se često pojavljuju u sižeima knjiga
- Stemming – Porter stemmer **NLTK**
- Lematizacija – **NLTK**

Pred-procesiranje korpusa 1/2

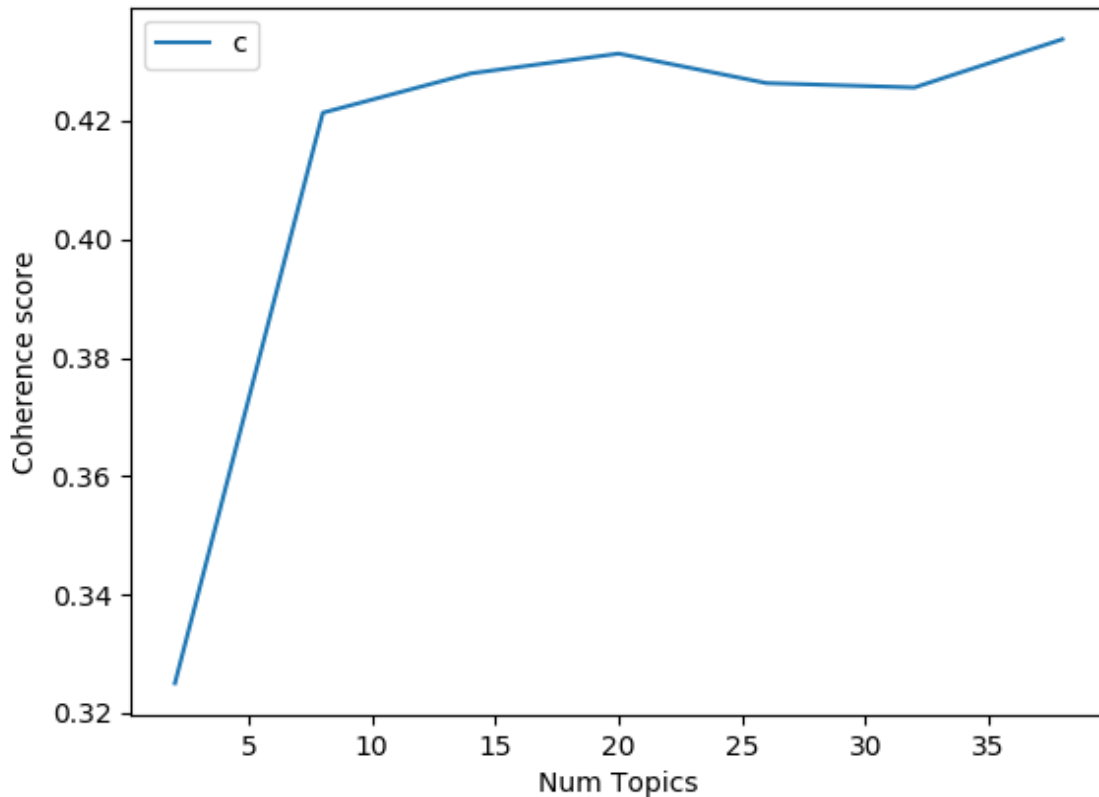
- Uklanjanje reči koje imaju manje 3 tri karaktera.
- Uklanjanje reči:
 - koje se pojavljuju u više od 50% dokumenata,
 - reči koje se ukupno javljaju manje od 15 puta.
- Nakon toga je ostavljeno 100.000 reči koje se najčešće pojavljuju dok su ostale takođe uklonjene iz korpusa.
- Upotrebom **Stanford NER** modela identifikovana su i uklonjena sva lična imena koja se pojavljuju u okviru korpusa

Određivanje optimalnog broja tema

- Lakat metoda kao kod određivanja optimalnog broja klastera
- Kao mera kvaliteta koristi se *koherentnost* modela
 - kreira se velik broj LDA modela sa različitim brojem za optimalan broj tema a zatim se za svaki od njih izračuna koherentnost, da bi se utvrdio najbolji model.
 - Zatim se ove vrednosti predstavljaju uz pomoć grafika a potom se prva vrednost za koju se javlja "lakat" na grafiku uzima kao optimalan broj tema.

Određivanje optimalnog broja tema

- Optimalan broj tema u ovom slučaju je 7 ili 8.



LDA implementacije

- Eksperimentisano je sa **Gensim** im **MALLET** alatima
- **MALLET** je odabran zato što pruža mogućnost primene modela
 - obučavanje na jednom korpusu
 - onda prepoznavanje tema iz obučavajućeg korpusa na drugom korpusu

Rezultati

Topic 0
army battle
order force
attempt power soldier
city
return plan

Topic 1
save magic
kill king power
travel
escape
attack return city

Topic 2
young woman father
family love
mother child
life year daughter

Topic 3
work chapter american
state world
narrator story
include character book

Topic 4
school friend book year
start story people
time begin
life

Topic 5
kill murder
death work case
body police doctor dead
find

Topic 6
human captain
crew attack planet
earth ship
force world race

Topic 7
home
house
return find
give talk leave
asks night room

Rezultati

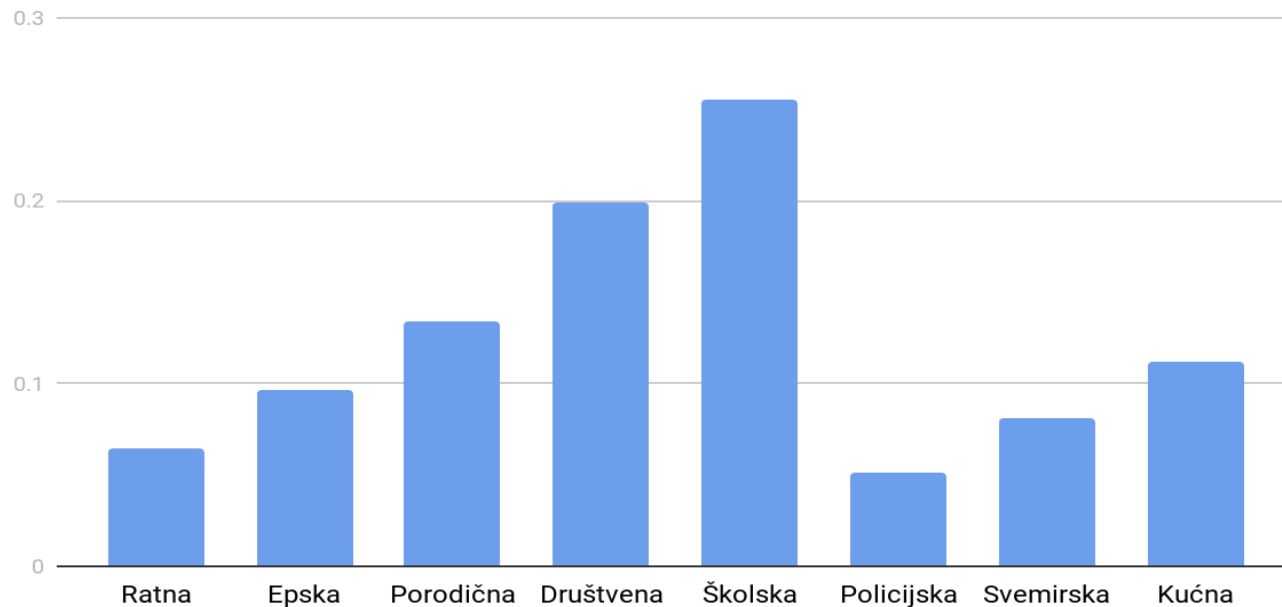
- Tumačenje tema od strane autora:

- Topic 0 - Ratna tema
- Topic 1 - Epska tema
- Topic 2 - Porodična tema
- Topic 3 - Društvena tema
- Topic 4 - Školska tema
- Topic 5 - Policijska tema
- Topic 6 - Svemirska tema
- Topic 7 - Kućna tema

Rezultati – primena na *Goodreads* korpus

- Jedan primer – „Lovac u žitu“

Verovatnoće tema



Slika 4.16: Verovatnoće tema u okviru sižea knjige *Lovac u žitu*