Основне академске студије
Информациони инжењеринг

Методе и технике науке о подацима

Увод у обраду природног језика и анализу текста

(материјали за предавања)

Садржај

Обрада природног језика и анализа текста

Обрада природног језика

енгл. *natural language processing*

рачунска лингвистика (енгл. *computational linguistics*)

потпоље рачунарских наука посвећено употреби рачунских техника за потребе учења, разумевања и стварања садржаја на људском језику (по Хиршберг и Менингу)

Обрада природног језика

- класификација дисциплина у вези с обрадом природног језика
  - класификација дисциплина у пољу рачунарства
    - *ACM Computing Classification System 2012*
    - https://dl.acm.org/ccs
  - обрада природног језика у контексту класификације *ACM Computing Classification System 2012*
    - *Методологије рачунарства*
      - *Вештачка интелигенција*
        - *Обрада природног језика*
          - *Издвајање информација*
          - *Машинско превођење*
          - *Дискурс, дијалог и прагматика*
          - *Генерисање природног језика*
          - *Препознавање говора*
          - *Лексичка семантика*
          - *Фонологија/морфологија*
          - *Језички ресурси*

Обрада природног језика и анализа текста

- актуелне теме и задаци
  - анализа сентимента (истраживање мишљења)
    - пример приступа
      - сентиментска природа целокупног текста као агрегација сентиментских природа појединачних речи
      - употреба сентиментских лексикона
        - за речи у лексикону могу бити назначени сентименти, уведене категорије или скорови за позитивни и негативни сентимент...
  - аутоматско превођење текста
    - пример приступа
      - примена рекурентних неуронских мрежа за превођење секвенце у секвенцу
        - мрежа типа кодер–декодер
  - генерисање текста
    - пример приступа
      - примена великих језичких модела заснованих на дубоком учењу
        - пример *GPT-4*
      - примена модела за конверзацију
        - пример *ChatGPT*
  - ....

Обрада природног језика и анализа текста

Основни концепти

- морфема
  - јединица која има значење
  - речи састављене од морфема
- стем
  - основна морфема у речи
  - даје главно значење речи
- афикс
  - додатна морфема у речи
  - уноси додатна значења у реч
- лема
  - скуп лексичких форми које су истог стема, исте главне врсте речи и истог значења речи
- токен
  - смислена јединица текста

Основни концепти

- стоп-реч
  - реч која се уклања из текста јер је иначе честа и обично нема велики допринос општем значењу текста
- униграм
  - појединачна реч из текста
- *n*-грам
  - секвенца *n* узастопних речи из текста
- корпус
  - машински читљива колекција текстова

Основни концепти

- нормализација текста
  - трансформација текста у стандардни (нормални) облик који је погодан за даљу употребу
  - поступак нормализације тесно повезан с језиком текста
  - често коришћене врсте поступака за нормализацију текста
    - токенизација речи (сегментација речи)
    - нормализација формата речи
    - сегментација реченица

Основни концепти

- токенизација
  - енгл. *tokenization*
  - растављање текста на токене
  - токен
    - секвенца (група) знакова
    - смислена јединица текста
      - знак, знак интерпункције, реч, *n*-грам, пасус, реченица...
      - токен је углавном реч
  - често коришћени облици токенизације
    - токенизација по речима
    - токенизација по знаковима
    - токенизација на основу података
      - формирање речника токена (скупа токена) на основу корпуса за обучавање и растављање циљног текста на токене који су доступни у формираном речнику

Основни концепти

- нормализација формата речи
  - трансформација речи или токена у стандардни (нормални) облик који је погодан за даљу употребу
  - често коришћене врсте нормализације
    - преклапање величине слова (енгл. *case folding*)
      - свођење свих слова на мала слова
      - углавном се примењује мада величина слова може бити корисни у одређеним контекстима (нпр. у анализи сентимента)
    - стемовање (енгл. *stemming*)
    - лематизација (енгл. *lemmatization*)

Основни концепти

- стемовање
  - свођење речи на њен стем
  - обично се своди на уклањање суфикса из речи

Основни концепти

- стемовање
  - Портеров алгоритам за стемовање (Портеров стемер)
    - радови Мартина Портера из 1979. и 1980. године
    - развијен за енглески језик
    - актуална каноничка верзија алгоритма дата у програмском језику *C* (*ANSI C)*
    - алгоритам за уклањање суфикса
    - заснива се на примени више правила
      - примери коришћених правила
        - уклањање наставака за множину
        - уклањање наставака *-ed*, *-ing*
        - замена двоструких наставака једноструким
          - *tional >> tion*
          - *izer >> ize*
          - *icate >> ic*
          - *ical >> ic*
          - ...
    - препорука да се користи стемер типа *Snowball*

Основни концепти

- лематизација
  - свођење речи на њену лему
  - лема очекивано налази се у речнику
  - лематизација начелно захтевнији поступак од стемовања

Основни концепти

- сегментација реченице
  - растављање текста на реченице
  - растављање се углавном изводи на основу знакова интерпункције којима се означава крај реченице
    - растављање углавном на основу тачке, упитника или узвичника
    - постоје одступања јер знакови интерпункције могу имати различите улоге у тексту

Основни концепти – пример

- текст
  - *I was unable to follow the whole argument, but it was evident that the English Professor had handled his subject in a very aggressive fashion, and had thoroughly annoyed his Continental colleagues.*
    - текст из дела *Изгубљени свет* од Артура Конана Дојла
- технологија
  - примена библиотеке *NLTK 3.8.1* за језик *Python 3.11.5*
- кораци
  - примена токенизатора типа *Punkt*
    - растављање по знаковима интерпункције (осим по тачкама)
  - уклањање токена који одговарају знаковима интерпункције
    - циљ да преостану токени који одговарају речима
  - обрада речи
    - примена Портеровог стемера
    - примена стемера типа *Snowball*
    - примена лематизатора типа *WordNet*

Основни концепти – пример

резултат 1/2

REČ	STEM PORTER	STEM SNOWBALL	LEMA WORDNET
I	i	i	I
was	wa	was	wa
unable	unabl	unabl	unable
to	to	to	to
follow	follow	follow	follow
the	the	the	the
whole	whole	whole	whole
argument	argument	argument	argument
but	but	but	but
it	it	it	it
was	wa	was	wa
evident	evid	evid	evident
that	that	that	that
the	the	the	the
English	english	english	English

...

Основни концепти – пример

резултат 2/2

...

Professor	professor	professor	Professor
had	had	had	had
handled	handl	handl	handled
his	hi	his	his
subject	subject	subject	subject
in	in	in	in
a	a	a	a
very	veri	veri	very
aggressive	aggress	aggress	aggressive
fashion	fashion	fashion	fashion
and	and	and	and
had	had	had	had
thoroughly	thoroughli	thorough	thoroughly
annoyed	annoy	annoy	annoyed
his	hi	his	his
Continental	continent	continent	Continental
colleagues	colleagu	colleagu	colleague

Примери софтверских технологија намењених обради природног језика

- *NLTK* (*Natural Language Toolkit*)
  - развој од 2001. године на Универзитету Пенсилваније (Филаделфија, Пенсилванија, САД) уз додатно учешће волонтера
  - технологија заснована на програмском језику *Python*
  - обухвата бројне корпусе
  - Интернет сајт
    - https://www.nltk.org/

Примери софтверских технологија намењених обради природног језика

- *CoreNLP*
  - рад Групе за обраду природног језика на Универзитету Стенфорд (Стенфорд, Калифорнија, САД)
  - технологија заснована на програмском језику *Java*
  - функционисање по моделу тока и примене разноврсних анотатора над текстом
  - подршка за арапски, енглески, италијански, кинески, мађарски, немачки, француски и шпански
  - Интернет сајт
    - https://stanfordnlp.github.io/CoreNLP/

Примери софтверских технологија намењених обради природног језика

- *spaCy*
  - рад компаније *Explosion*
  - технологија заснована на програмском језику *Cython*
  - усмереност на примену у пракси
  - подршка за велики број језика
  - Интернет сајт
    - https://spacy.io/

Пример корпуса

- корпус *Penn Treebank*
  - као резултат пројекта формиран је велики анотирани корпус за енглески језик америчког стандарда
    - обухвата преко 4,5 милиона речи
  - верзија *Treebank-3* из 1999. године
    - https://catalog.ldc.upenn.edu/LDC99T42

Пример лексичке базе података

- база података *WordNet*
  - лексичка база података за енглески језик
  - историјски развој
    - почетак
      - рад групе психолога и лингвиста на Универзитету Принстон (Принстон, Њу Џерси, САД) од 1985. године
    - најновија верзија 3.1
      - https://wordnet.princeton.edu/download/current-version
  - формат
    - више датотека у формату *ASCII*
      - поља углавном раздвојена размаком
      - редови се завршавају знаком за нови ред

Пример лексичке базе података

- база података *WordNet*
  - обухваћене врсте речи
    - именице, глаголи, придеви и прилози
  - груписање речи у синскупове који су повезани разним врстама веза
    - синскуп као неуређени скуп когнитивних синонима
    - лексичке везе
      - између семантички повезаних облика речи
    - семантичке везе
      - између значења речи

Пример лексичке базе података

- база података *WordNet*
  - везе
    - синонимија
      - синоними означавају исти концепт
    - антонимија
      - антоними означавају супротстављене концепте
    - хипернимија и хипонимија
      - однос надређености и подређености
        - однос типа *ISA*
      - хиперним представља класу посебних инстанци
      - хипоним представља припадника неке класе
    - холонимија и меронимија
      - однос типа целина–део
      - холоним представља целину за неки део
      - мероним представља део за неку целину
    - тропонимија
      - однос разрађивања
      - тропоним представља посебан начин разрађивања
    - ...

Садржај

Извори и литература

Основни извори и литература

- Hirschberg J, Manning J. Advances in Natural Language Processing. Science. 2015 July;349(6245); 261–266.
- Jurafsky D, Martin JH. Speech and Language Processing. 3rd edition draft. 2025. Internet: https://web.stanford.edu/~jurafsky/slp3/
- Bird S, Klein E, Loper E. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. 2019. Internet: https://www.nltk.org/book/
- Silge J, Robinson D. Text Mining with R. O'Reilly (Sebastopol, CA, USA); 2017. Internet: https://www.tidytextmining.com/index.html
- Géron A. Mašinsko učenje: Scikit-Learn, Keras i TensorFlow: koncepti, alati i tehnike za izgradnju inteligentnih sistema. Prevod 2. izdanja. O'Reilly (Sebastopol, CA, USA), Mikro knjiga (Beograd, Srbija); 2021.

Извори и литература

Основни извори и литература

- Tartarus.org. The Porter Stemming Algorithm. Internet: https://tartarus.org/martin/PorterStemmer/
- Snowball. Snowball. Internet: https://snowballstem.org/
- ACM Digital Library. Computing Classification System. Internet: https://dl.acm.org/ccs
- OpenAI. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Internet: https://openai.com/index/gpt-4/
- OpenAI. GPT-4. Internet: https://openai.com/index/gpt-4-research/
- OpenAI. GPT-4 Technical Report. 2023. arXiv:2303.08774. Internet: https://arxiv.org/abs/2303.08774
- OpenAI. Introducing ChatGPT. Internet: https://openai.com/index/chatgpt/

Извори и литература

Основни извори и литература

- Project Gutenberg. The Lost World by Arthur Conan Doyle. Internet: https://www.gutenberg.org/ebooks/139
- NLTK Project. Natural Language Toolkit. Internet: https://www.nltk.org/
- NLTK Project. nltk package. Internet: https://www.nltk.org/api/nltk.html
- nltk. FAQ · nltk/nltk Wiki · GitHub. Internet: https://github.com/nltk/nltk/wiki/FAQ
- CoreNLP. CoreNLP. Internet: https://stanfordnlp.github.io/CoreNLP/
- Stanford NLP Group. The Stanford Natural Language Processing Group. Internet: https://nlp.stanford.edu/
- spaCy. Industrial-strength natural language processing. Internet: https://spacy.io/

Извори и литература

Основни извори и литература

- Marcus MP, Santorini B, Marcinkiewicz MA. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics. 1993 June;19(2); 313–330.
- Linguistic Data Consortium. Treebank-3. Internet: https://catalog.ldc.upenn.edu/LDC99T42
- WordNet. WordNet. Internet: https://wordnet.princeton.edu/
- WordNet. WordNet publications. Internet: https://wordnet.princeton.edu/publications
- WordNet. Current version. Internet: https://wordnet.princeton.edu/download/current-version
- WordNet. WordNet documentation. Internet: https://wordnet.princeton.edu/documentation
- WordNet. Frequently asked questions. Internet: https://wordnet.princeton.edu/frequently-asked-questions

Извори и литература

Основни извори и литература

- WordNet. wnintro(5WN). Internet: https://wordnet.princeton.edu/documentation/wnintro5wn
- WordNet. wngloss(7WN). Internet: https://wordnet.princeton.edu/documentation/wngloss7wn

Извори и литература

Додатни извори и литература

- WordNet Search. WordNet Search - 3.1. Internet: http://wordnetweb.princeton.edu/perl/webwn
- TensorFlow. Embedding projector. Internet: https://projector.tensorflow.org/

Основне академске студије
Информациони инжењеринг

Методе и технике науке о подацима

Увод у обраду природног језика и анализу текста

(материјали за предавања)