Szkoła Letnia - Krakow Applied Physics and Computer Science Summer School

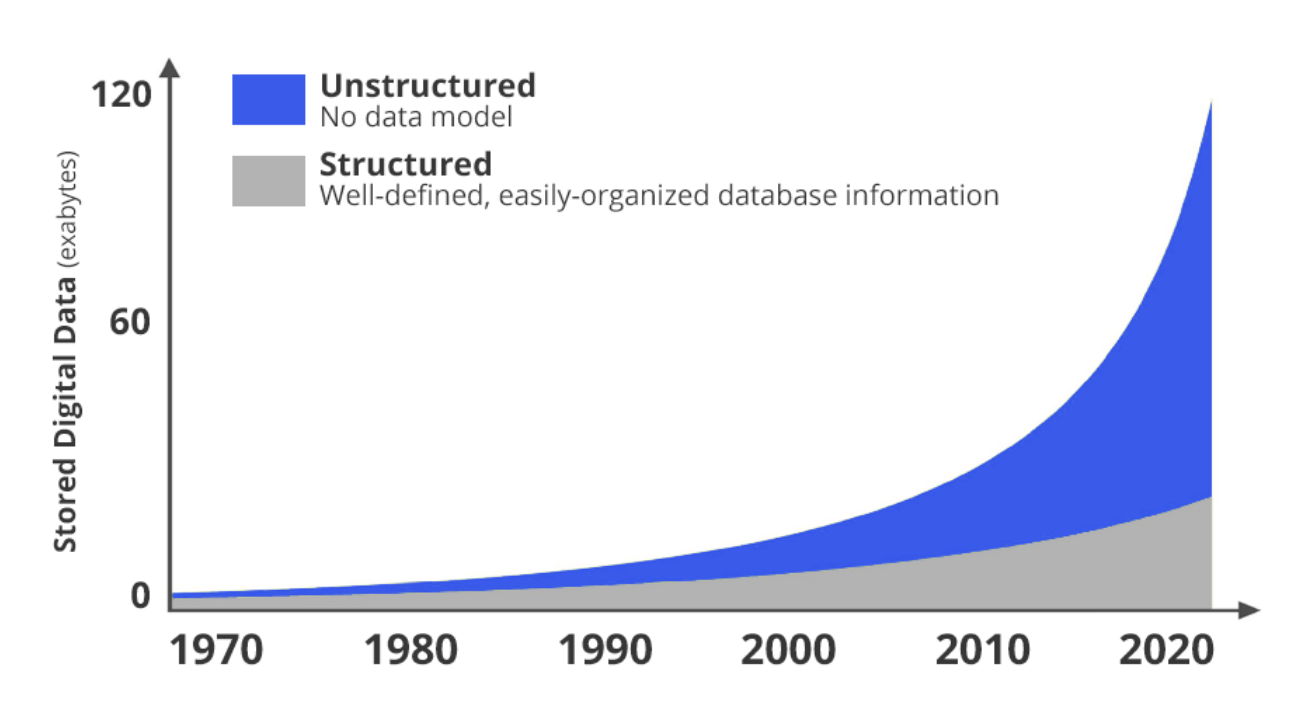
Wybrane zastosowania uczenia maszynowego w przetwarzaniu obrazów i języka naturalnego, cz. 1

Szymon Łukasik

Wydział Fizyki i Informatyki stosowanej AGH

Dlaczego taki temat?

Według wielu badań (m.in. Oracle) – zdecydowana większość danych ma postać nieustrukturyzowaną (obrazy, teksty, filmy itp.).



http://idc.com

Co łączy te algorytmy?

K-średnich

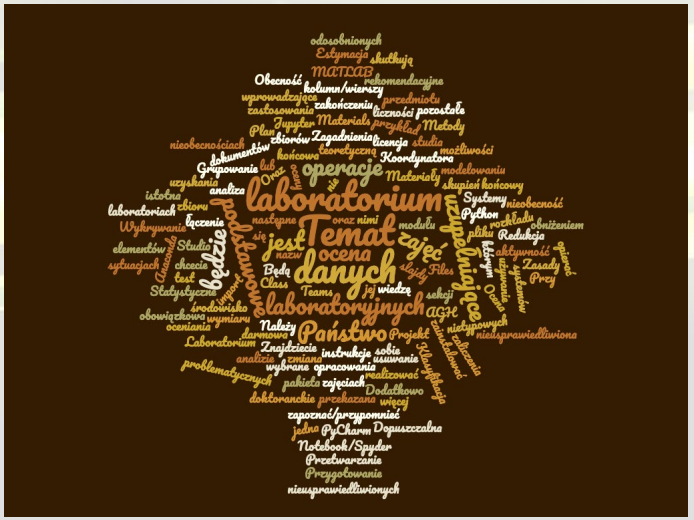
3-sigma

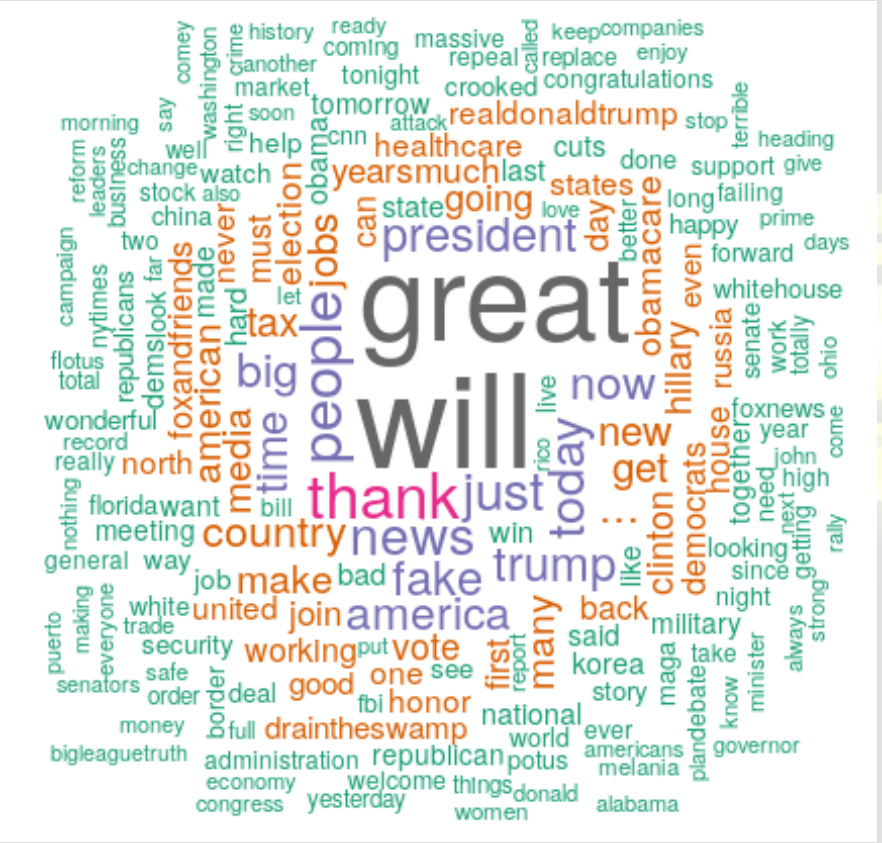
K-najbliższych sąsiadów

Czasami liczby nie są liczbami

Zacznijmy od tekstu

Eksploracja tekstu /text mining/ zwana również eksploracją danych tekstowych, w przybliżeniu równoważna analizie tekstu, to proces uzyskiwania wysokiej jakości informacji z tekstu.





Źródła danych tekstowych

- strony internetowe
- wiadomości e-mail
- dane medyczne
- ankiety
- raporty o incydentach
- ulotki leków
- wiadomości
- mniej oczywiste: opisty produktów, koszyki, sekwencje DNA

Typowe problemy analizy tekstów

1. Analiza pojedynczego dokumentu

Informacje uzyskuje się z treści pojedynczego dokumentu, np. analizy sentymentu, generowania podsumowań.

2. Analiza zestawów dokumentów

Wydobywanie informacji opiera się na zbiorze tekstów - np. poprzez ich kategoryzację, wykrywanie niestandardowych tekstów, atrybucję itp.

Ilościowa reprezentacja dokumentów

Najprostsza reprezentacja – binarna macierz występowania słów

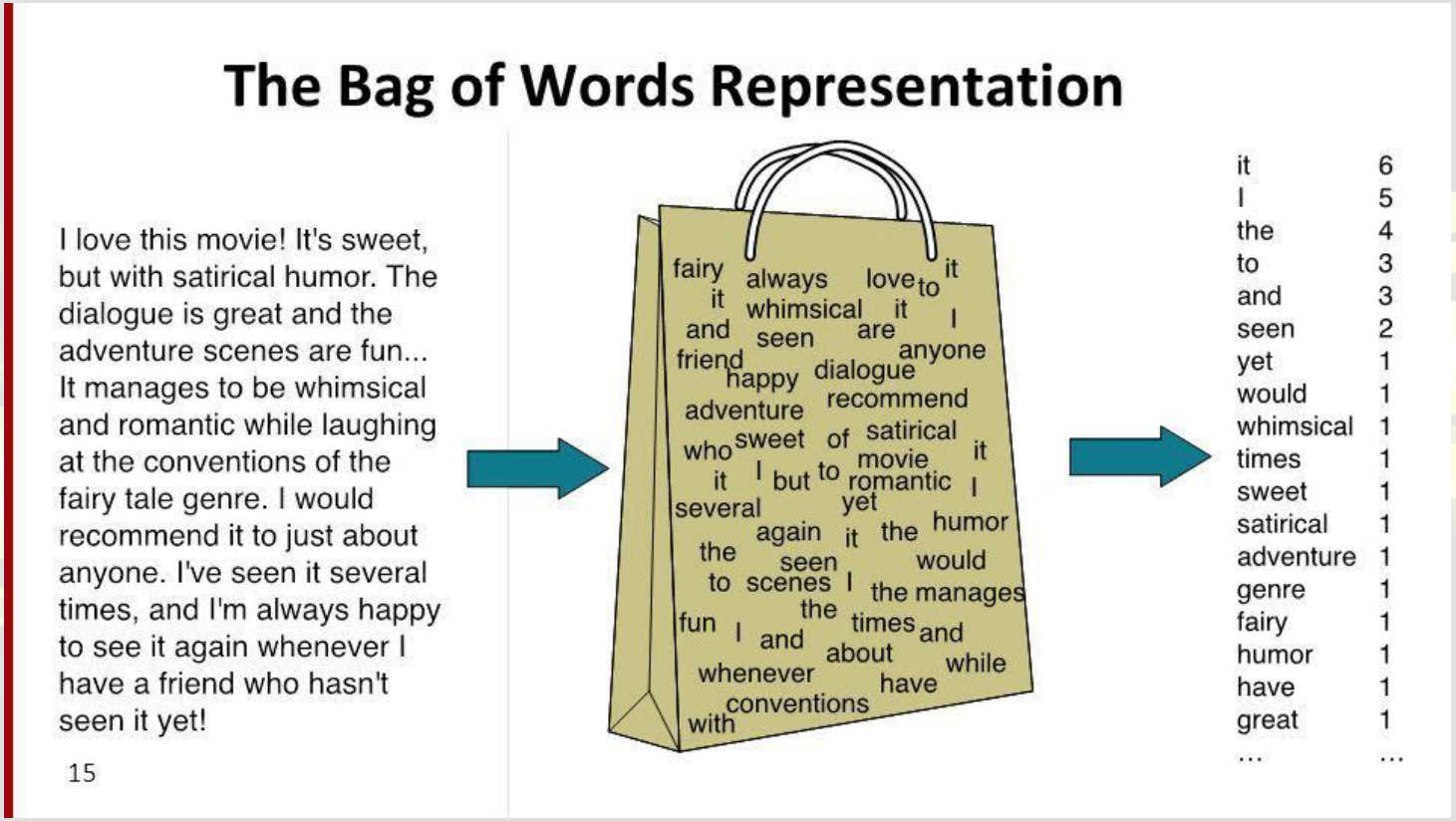
	Chili	Chair	Mat	Mouse	Bike	Toy
D1	1	1	1	0	0	1
D2	1	1	1	0	1	0
D3	1	1	1	0	0	0
D4	1	1	1	0	0	0
D5	1	1	1	0	1	0
D6	1	0	0	1	1	1
D7	0	0	1	1	1	0
D8	1	0	0	1	1	1
D9	1	0	0	1	1	1
D10	1	0	0	1	1	1

Ilościowa reprezentacja dokumentów

Tą reprezentację można poszerzyć do term frequency (TF) – powszechnie znany jako bag-of-words.

	Chili	Chair	Mat	Mouse	Bike	Toy
D1	24	21	9	0	0	3
D2	32	10	5	0	3	0
D3	12	16	5	0	0	0
D4	6	7	2	0	0	0
D5	43	31	20	0	3	0
D6	2	0	0	18	7	6
D7	0	0	1	32	12	0
D8	3	0	0	22	4	4
D9	1	0	0	34	27	25
D10	6	0	0	17	4	23

Bag of Words



https://www.programmersought.com/article/4304366575/

Reprezentacja TF-IDF

Term-Frequency – Inverse Document Frequency (TF-IDF) dla słowa i w dokumencie j jest dany jako:

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_i}$$

gdzie:

tf_{ij} – liczba wystąpień słowa i w dokumencie j

N – liczba dokumentów

df_i – liczba dokumentów zawierających słowo i

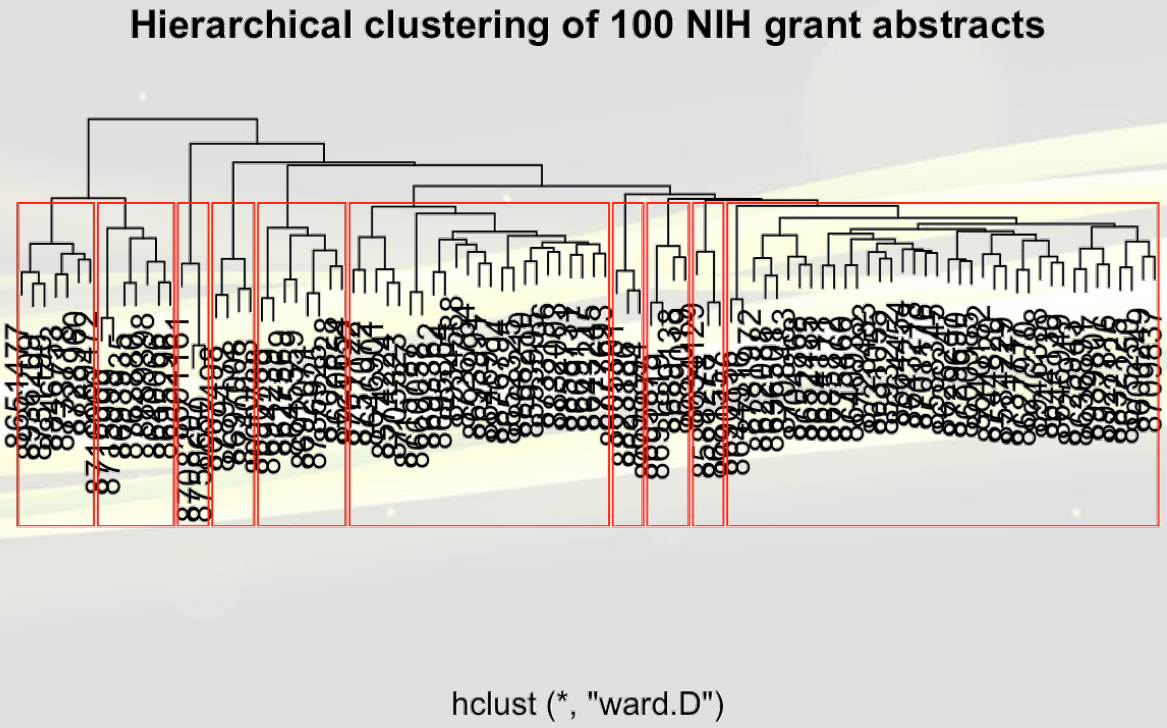
Wstępne przetwarzanie dokumentu

1. Usuwanie znaków interpunkcyjnych z każdego przetwarzanego dokumentu i tworzenie, niezależnie dla każdego przetworzonego tekstu, listy słów w nim obecnych
2. Usunięcie nieistotnych słów z punktu widzenia dalszej analizy
3. Konwertowanie wyrazów pojawiających się na listach na ich podstawową formę. Transformacja do formy podstawowej jest również znana jako stemming. Zwykle realizowane za pomocą reguł lub za pomocą słowników.

4. https://pypi.org/project/pystempel/

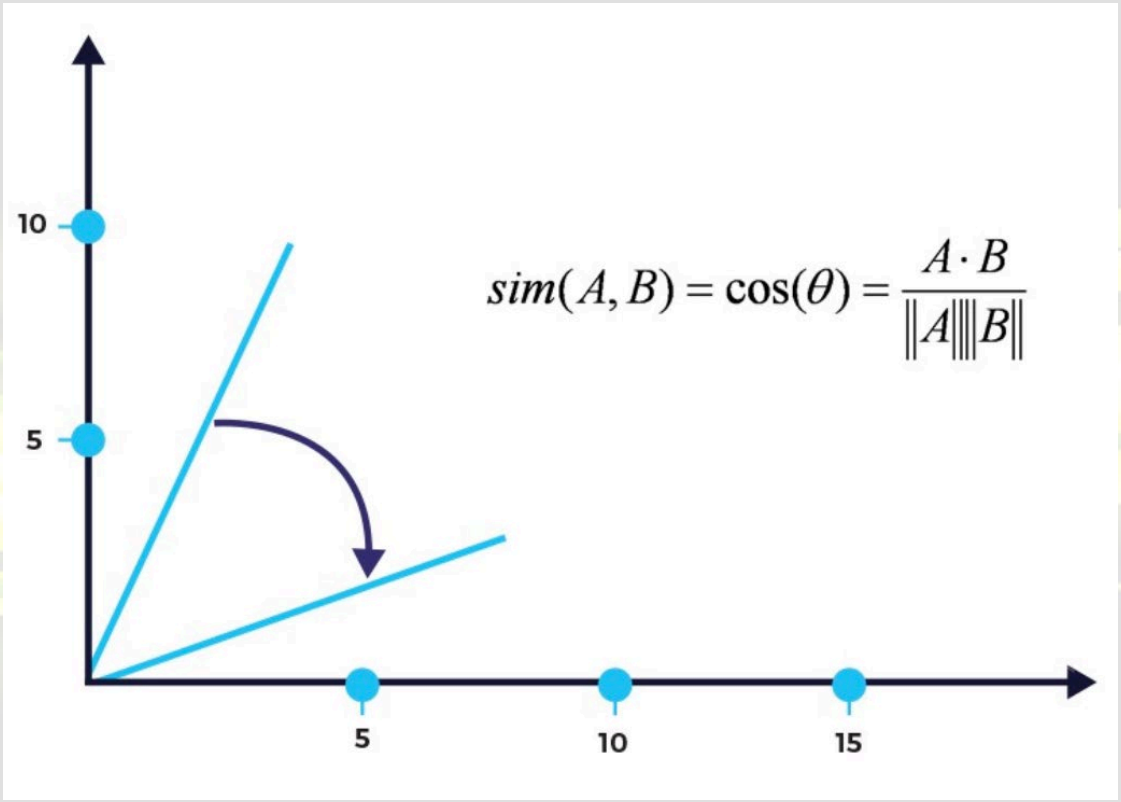
```
>>> for word in ['książka', 'książki', 'książkami', 'książkowa', 'książkowymi']:
...   print(stemmer.stem(word))
...
książek
książek
książek
książkowy
książkowy
```

Przykład



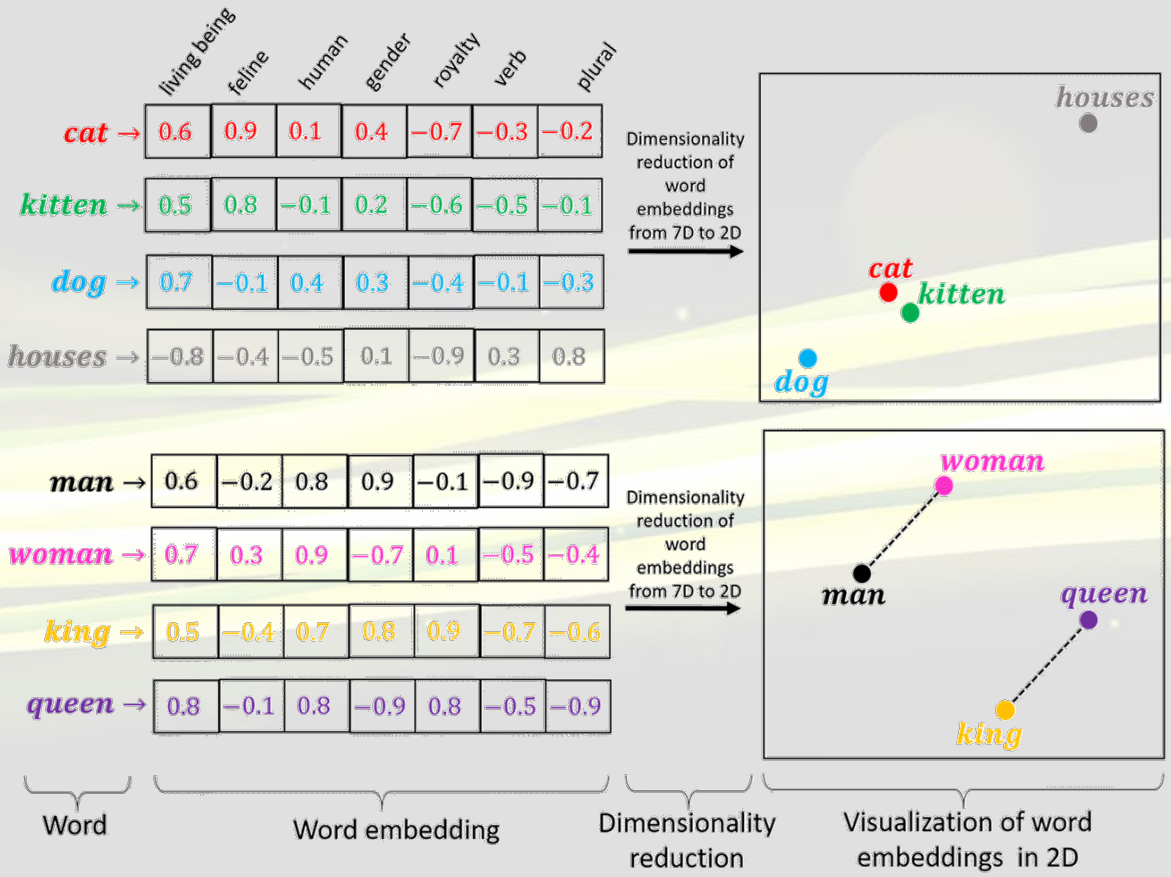
https://www.rtextminer.com/articles/b_document_clustering.html

Podobieństwo?



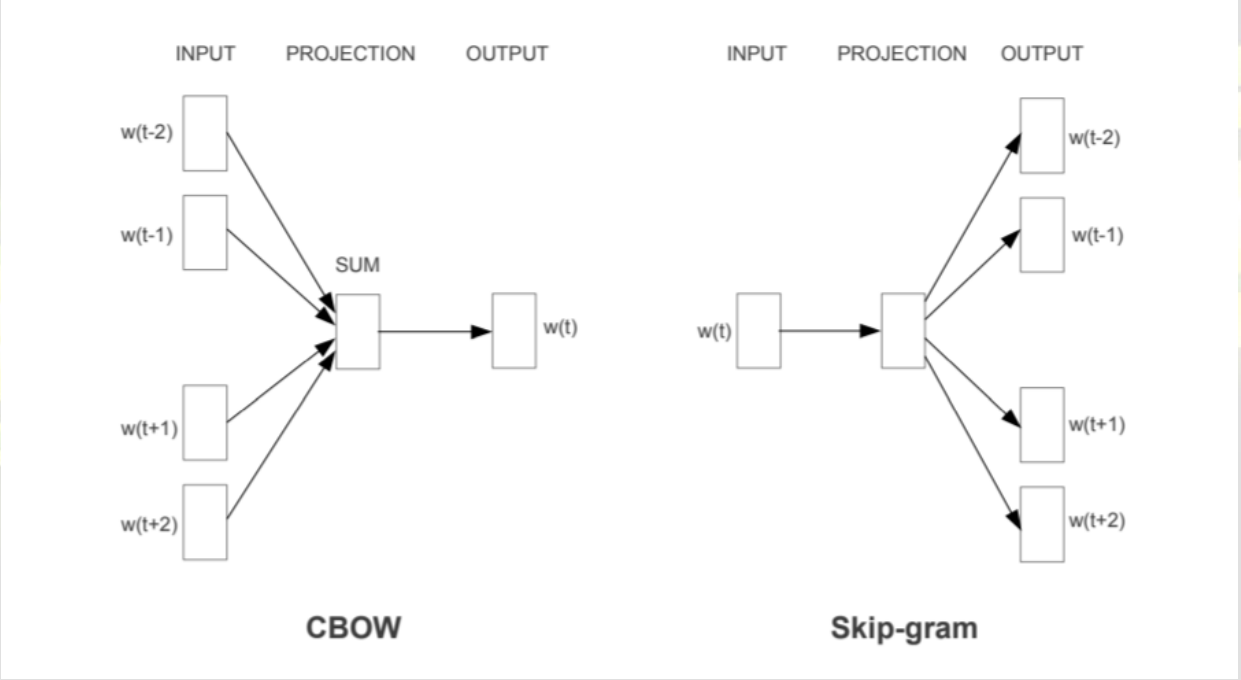
https://intellica.ai/

Embeddingi słów



Embeddingi słów

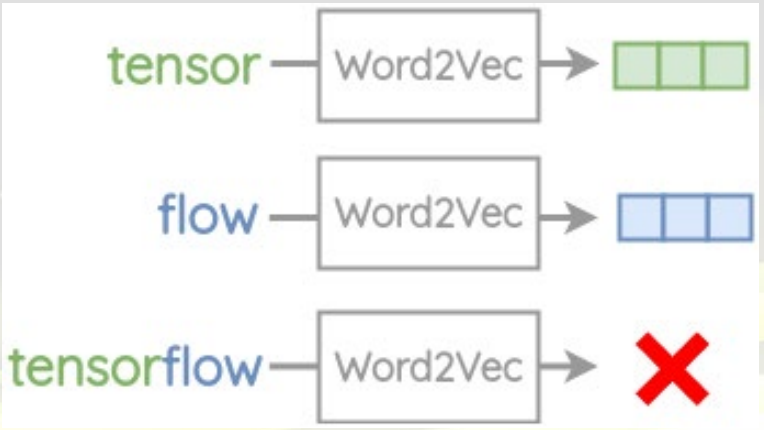
Pierwsze podejście – word2vec

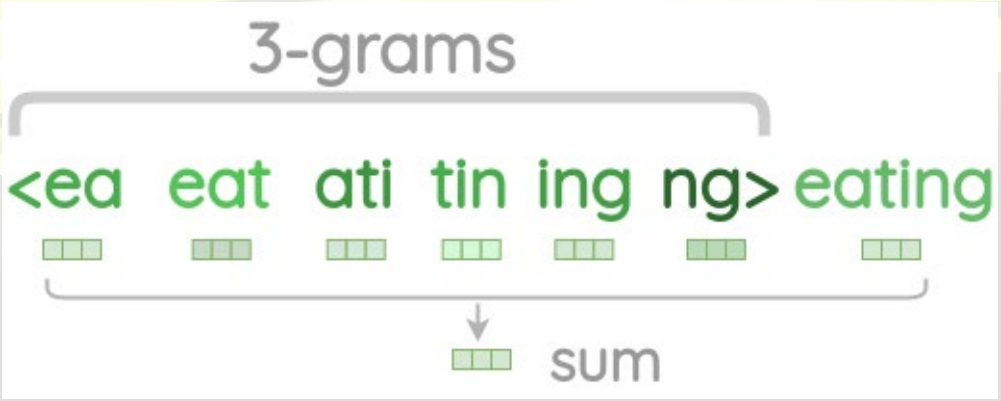


Skip-gram

Source Text	Training Samples generated from source text				
I will have orange juice and eggs for breakfast	(will, I)	(will, have)	(will, orange)		
I will have orange juice and eggs for breakfast	(have, I)	(have, will)	(have, orange)	(have, juice)	
I will have orange juice and eggs for breakfast	(orange, will)	(orange, have)	(orange, juice)	(orange, and)	
I will have orange juice and eggs for breakfast	(juice, have)	(juice, orange)	(juice, and)	(juice, eggs)	
I will have orange juice and eggs for breakfast	(and, orange)	(and, juice)	(and, eggs)	(and, for)	
I will have orange juice and eggs for breakfast	(eggs, juice)	(eggs, and)	(eggs, for)	(eggs, breakfast)	
I will have orange juice and eggs for breakfast	(for, and)	(for, eggs)	(for, breakfast)		

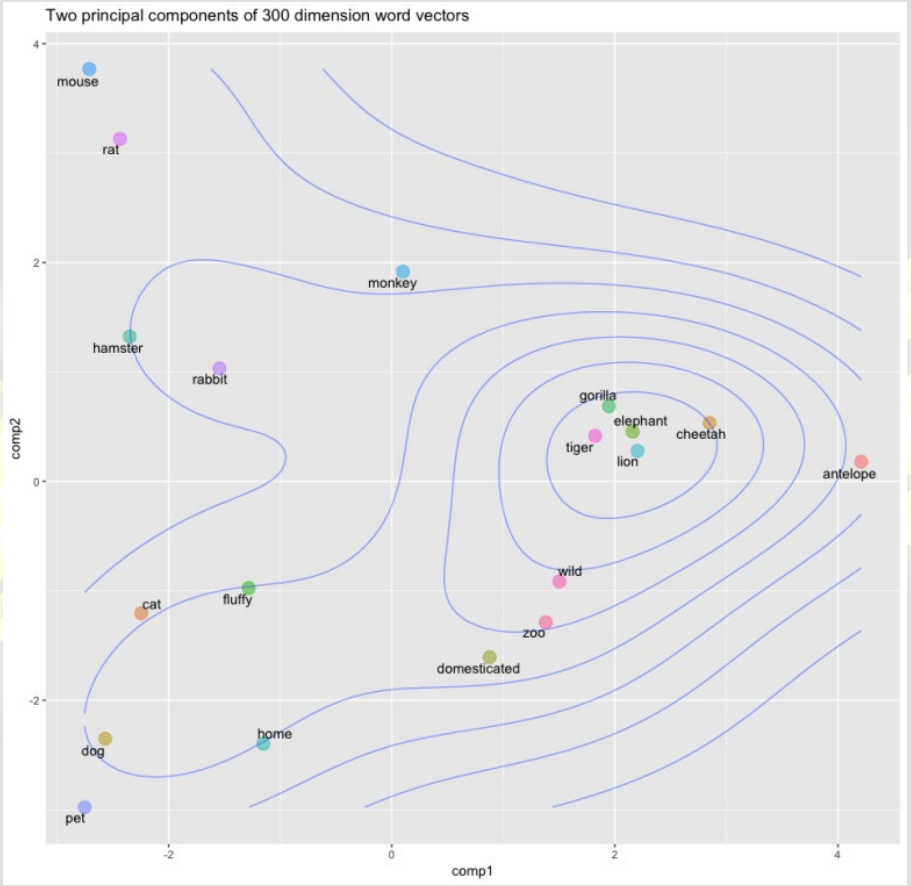
FastText – n-gramy





https://amitness.com/2020/06/fasttext-embeddings/

Przykład


Two principal components of 300 dimension word vectors
mouse
rat
monkey
hamster
rabbit
gorilla
elephant
cheetah
tiger
lion
antelope
wild
cat
fluffy
zoo
domesticated
home
dog
pet
comp1
comp2
4
2
0
-2
-2
0
2
4

W praktyce

https://spacy.io/

https://fasttext.cc/

https://radimrehurek.com/gensim/

https://github.com/google-research/bert

Przykład

Build a Multi-Layer Perceptron

We are now ready to build our model. We will use an Embedding layer to map from an integer that corresponds to a word, to a vector of floating point weights (the embedding). These weights are learned when we train the model.

```
from keras.models import Sequential
from keras.layers import Dense, Embedding, GlobalAveragePooling1D

embedding_dimension = 16

model = Sequential()
model.add(Embedding(num_words, embedding_dimension, input_length=max_len))

# Our output is a 3D tensor of shape (samples, vocab_size, embedding_dimension)
# we will use `GlobalAveragePooling` before our fully connected layer. In the past, this used to be a Flatten layer
model.add(GlobalAveragePooling1D())

# Add a classifier on top.
model.add(Dense(1, activation='sigmoid') )
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])

model.summary()

history = model.fit(
    train_data,
    train_labels,
    epochs=10,
    batch_size=32,
    validation_split=0.2
)
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 500, 16)           320000
_________________________________________________________________
global_average_pooling1d_1 ( (None, 16)                0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 17
```

SPEECH AND TEXT PROCESSING APIs COMPARISON

	Amazon	Microsoft	Google	IBM
Speech Recognition (Speech into Text)	✓	✓	✓	✓
Text into Speech Conversion	✓	✓	✓	✓
Entities Extraction	✓	✓	✓	✓
Key Phrase Extraction	✓	✓	✓	✓
Language Recognition	100+ languages	120 languages	120+ languages	60+ languages
Topics Extraction	✓	✓	✓	✓
Spell Check	✗	✓	✗	✗
Autocompletion	✗	✓	✗	✗
Voice Verification	✓	✓	✗	✗
Intention Analysis	✓	✓	✓	✓
Metadata Extraction	✗	✗	✗	✓
Relations Analysis	✗	✓	✗	✓
Sentiment Analysis	✓	✓	✓	✓
Personality Analysis	✗	✗	✗	✓
Syntax Analysis	✗	✓	✓	✓
Tagging Parts of Speech	✗	✓	✓	✗
Filtering Inappropriate Content	✗	✓	✓	✗
Low-quality Audio Handling	✓	✓	✓	✓
Translation	6 languages	60+ languages	100+ languages	48 languages
Chatbot Toolset	✓	✓	✓	✓

altexsoft
software r&d engineering

Comparison of the APIs for Speech Processing

	API	Tasks supported	Main details	Languages supported	Results quality
amazon	Transcribe	Speech to text converting	Punctuation and formatting, telephony audio, customization and multiple speakers recognition	English Spanish	GOOD
	Polly	Text to speech converting	Real-time mode, pronunciation, volume, pitch, speed rate, etc customization	27 + dialects	EXCELLENT
Google Cloud	Speech API	Speech to text converting	Customization, batch and real-time modes, noise robustness, filters for wrong words relative to the context, flexibility in the source files storage	120	INTERMEDIATE
IBM Watson	Speech to Text	Speech to text converting	Real-time mode, custom models, keywords spotting, speaker labels (in beta), word confidence, word timestamps, profanity filtering, word alternatives, smart formatting (in beta)	11	GOOD
	Text to Speech	Text to speech converting	Pronunciation customization, custom words, expressiveness, word timings	8 + dialects	EXCELLENT
Microsoft Azure	Bing Speech API	Speech to text converting	Real-time mode, customization, formatting, profanity filtering, text normalization, integration with Azure LUIS, speech scenarios	10 conversational mode 29 + dialects interactive and dictation modes	GOOD
		Text to speech converting	Pronunciation, volume, pitch etc customization	78 + dialects	EXCELLENT
nexmo	Voice API	Text to speech converting	Different genders, accents	23	-
SPEECHMATICS	ASR	Speech to text converting	Real-time mode, specialized on English (Global English), sentences boundaries, words timing, confidences	75	INTERMEDIATE
twilio	Speech Recognition	Speech to text converting	Real-time mode, profanity filter	119 + dialects	-
VOCAPIA	Sigma API	Speech to text converting	Real-time mode, speaker labels, word timings, confidences, punctuations, language identification tags, specific entities recognition, customization	17	-

Created by ActiveWizards

A samemu?

Title	Korpus nagrań próbek mowy do celów budowy modeli akustycznych dla automatycznego rozpoznawania mowy w języku polskim, cz. 1
Persons	**Authors:** Teresa Sas **Partner:** Wrocław University of Science and Technology
Description	Korpus nagrań próbek mowy do celów budowy modeli akustycznych dla automatycznego rozpoznawania mowy w języku polskim. Jest to zbiór nagrań zapisanych w postaci standardowych plików audio, uzupełniony o ich transkrypcję tekstową. Zbiór został odpowiednio usystematyzowany i skatalogowany, oznaczona została płeć mówcy. Podzielony został na nagrania wykonane w trybie nadzorowanym, gdzie mówca czytał dostarczony mu tekst oraz nagrania nienadzorowane, spontaniczne, zapisy audio wywiadów i wystąpienia konferencyjne pracowników naukowych rejestrowane na żywo. Dodatkowo oznaczone zostały tak zwane nagrania zakłócone, czyli nagrania wykonane w trudnych warunkach akustycznych, ze słyszalnym pogłosem, muzyką w tle lub dużą ilością artefaktów typu y, yyy, e, eee, m, mmm, hmh. Tak przygotowany korpus może być wykorzystany zarówno do budowy modeli akustycznych dla systemów rozpoznawania mowy dla języka polskiego jak i stanowić bazę do prowadzenia badań naukowych z zakresu sztucznej inteligencji, rozumienia mowy, uczenia maszynowego, sieci głębokiego uczenia i wielu innych zagadnień naukowych związanych z inżynierią języka naturalnego. (Polish) **Comments:** W pliku pdf opisano szczegółowo sposób tworzenia korpusu oraz podano listę dokumentów, które w nim uwzględniono.
Keywords	"model"@pl, "korpus nagrań próbek mowy"@pl, "model akustyczny"@pl, "automatyczne rozpoznawanie mowy"@pl, "rozpoznawanie mowy"@pl, "rozumienie języka naturalnego"@pl, "anotacja korpusu"@pl
Classification	**Resource type:** dataset, database **Scientific discipline:** Dziedzina nauk inżynieryjno-technicznych / informatyka techniczna i telekomunikacja (2018) **Destination group:** general public, public administration, entrepreneurs, pupils, students, teachers, scientists **Harmful content:** No
Characteristics	**Place of creation:** Politechnika Wrocławska **Creation time:** 2020 **Resource language:** Polish
License	CC BY-SA 4.0
Technical information	**Submitter:** Teresa Sas **Availability date:** 16-11-2020
Collections	Kolekcja Politechniki Wrocławskiej

Koniec
– zapraszam jutro na troche obrazków