Przetwarzanie języka naturalnego

dr inż. Marcin Ciura

Wydział Informatyki i Telekomunikacji Politechniki Krakowskiej

Plan na dziś: 16 slajdów

- Dygresja: rekomendacje oparte na preferencjach użytkowników
- Rekomendacje oparte na treści

Dygresja: rekomendacje oparte na preferencjach użytkowników

Rekomendacje oparte na preferencjach użytkowników

Rekomendacje oparte na preferencjach użytkowników (*collaborative filtering*)

Collaborative filtering (1)

Proszę zajrzeć na slajdy 3. i 4. w pliku wyklad06b.pdf, których nie udało mi się tutaj przenieść

Collaborative filtering (2)

Wypełnij puste elementy macierzy M, rozkładając ją na iloczyn macierzy P i Q^T danego rzędu ($rank$) r tak, by zminimalizować błąd przybliżenia $\|M - PQ^T\|$ obliczony dla niezerowych elementów macierzy M

Collaborative filtering (3)

Uwagi praktyczne:

- W Pythonie:
  - `pip install implicit`
  - `implicit.als.AlternatingLeastSquares`
- Honorować ograniczenia, np. terytorialne lub wiekowe
- Obniżać punkty rekomendacjom, które były wcześniej pokazywane danemu użytkownikowi

Rekomendacje oparte na treści

Content-based filtering

1. Usuwanie słów nieinformatywmnych (*stop words*)
2. Stemming
3. Przekształcenie tekstu na multizbiór wyrazów (*bag of words*)
4. Przekształcenie TF–IDF
5. Szukanie najbliższych sąsiadów poprzez podobieństwo kosinusowe (*cosine similarity*)

Content-based filtering: wejście

we're going to start with two sticks of unsalted butter and to that i'm going to add four to five cups of confectioner's sugar depend skilful tins of milk to the butter and confectioner's sugar mixed or that'll do to make it look like a thing and to make a tastes fantastic brigand identities with the want vanilla extract flex we want to make the fasting pink i'm going to add a couple drops of liquid as for colorings of pink gel died that couldn't work really while tail and was mixed up together with the frosting until her no more color straightness all combined to nathan pink and pretty now and across the cub cakes for each cup cake and

Content-based filtering: usuwanie *stop words*

we're going to start with two sticks of unsalted butter and to that i'm going to add four to five cups of confectioner's sugar depend skilful tins of milk to the butter and confectioner's sugar mixed or that'll do to make it look like a thing and to make a tastes fantastic brigand identities with the want vanilla extract flex we want to make the fasting pink i'm going to add a couple drops of liquid as for colorings of pink gel died that couldn't work really while tail and was mixed up together with the frosting until her no more color straightness all combined to nathan pink and pretty now and across the cub cakes for each cup cake and

Content-based filtering: stemming

going start two sticks unsalted butter going add four five cups confectioner's sugar depend skilful tins milk butter confectioner's sugar mixed make look like thing make tastes fantastic brigand identities want vanilla extract flex want make fasting pink going add couple drops liquid colorings pink gel died work really tail mixed together frosting color straightness combined nathan pink pretty now across cub cakes cup cake

Content-based filtering: stemming

```
      go       start      two stick     unsalt
butter                 go       add four    five cup
   confectioner   sugar depend skilful tin     milk
       butter     confectioner   sugar mix
              make    look like   thing         make
taste  fantast   brigand identiti             want
vanilla extract flex    want    make     fast    pink
    go       add   couple drop     liquid
color        pink gel die                work real
      tail         mix      together
frost                     color straight
combin      nathan pink     pretty now     across
cub cake          cup cake
```

Content-based filtering: *bag of words*

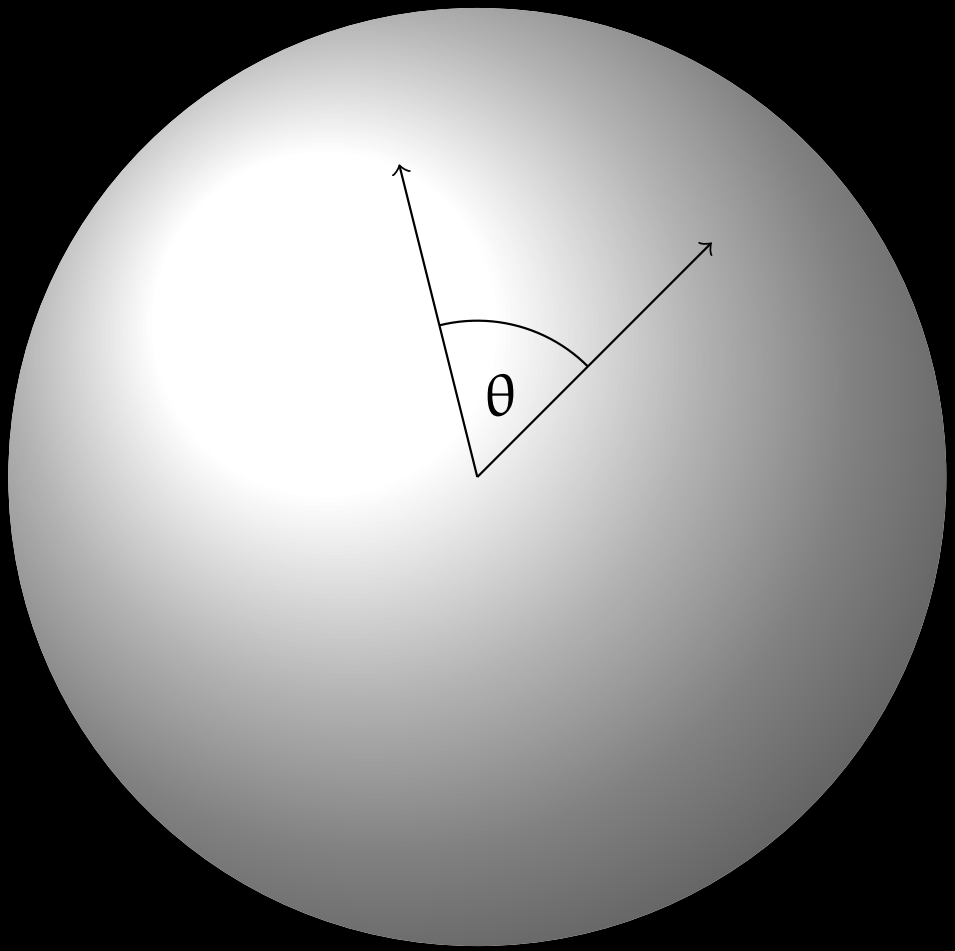
```
go: 3 make: 3 pink: 3 add: 3 butter: 2 cake: 2
color: 2 confectioner: 2 cup: 2 mix: 2 sugar: 2
want: 2 across: 1 brigand: 1 combin: 1 couple: 1
cub: 1 depend: 1 die: 1 drop: 1 extract: 1 fantast: 1
fast: 1 five: 1 flex: 1 four: 1 frost: 1 gel: 1
identiti: 1 like: 1 liquid: 1 look: 1 milk: 1
nathan: 1 now: 1 pretty: 1 real: 1 skilful: 1
start: 1 stick: 1 straight: 1 tail: 1 taste: 1
thing: 1 tin: 1 together: 1 two: 1 unsalt: 1
vanilla: 1 work: 1
```

Content-based filtering: TF-IDF

Term Frequency–Inverse Document Frequency
Obniża punkty często występujących wyrazów

```
go: 1.2 make: 1.4 pink: 2.6 add: 2.0 butter: 1.8
cake: 1.7 color: 1.6 confectioner: 1.8 cup: 1.6
mix: 1.2 sugar: 1.3 want: 0.9 across: 0.8
brigand: 1.0 combin: 0.9 couple: 0.8 cub: 1.0
depend: 0.7 die: 0.9 drop: 0.6 extract: 0.7
fantast: 0.8 fast: 0.7 five: 0.8 flex: 1.0 four: 0.8
frost: 1.0 gel: 1.0 identiti: 0.9 like: 0.6
liquid: 1.0 look: 0.6 milk: 0.8 nathan: 1.0 now: 0.6
pretty: 0.5 real: 0.5 skilful: 0.8 start: 0.7
stick: 0.9 straight: 0.9 tail: 0.8 taste: 0.8
thing: 0.5 tin: 1.0 together: 0.5 two: 0.8
unsalt: 1.0 vanilla: 1.0 work: 0.6 …
```

Content-based filtering: *cosine similarity*



$\cos 0° = 1$

$\cos 90° = 0$

$\cos 180° = -1$

Content-based filtering: *locality-sensitive hashing*

Przy milionach dokumentów przyspiesza wyszukiwanie podobnych dokumentów w stosunku do metody siłowej, czyli pętli po wszystkich dokumentach

Podsumowanie

- Dygresja: rekomendacje oparte na preferencjach użytkowników
- Rekomendacje oparte na treści

Do zobaczenia
na następnym wykładzie

o ???