SPRAWOZDANIE IV

METODY OBLICZENIOWE W NAUCE I TECHNICE

SINGULAR VALUE DECOMPOSITION

PROSTA WYSZUKIWARKA

DAWID BIAŁKA

2019/2020

Zadanie 1 Wyszukiwarka

Do wyszukiwarki możemy przesłać zapytanie długości od 2 do 5 słów i liczbę dokumentów, które chcemy, aby zostały wyświetlone. W odpowiedzi dostajemy podaną liczbę dokumentów w kolejności od najbardziej pasującego do danego zapytania. Kod wyszukiwarki znajduje się w pliku '1.py'.

Bazą dokumentów pod wyszukiwarkę jest nieco ponad 2 tysiące wpisów z angielskiego bloga o II wojnie światowej http://ww2today.com/. Wpisy te zostały pobrane przy pomocy web crawlera (framework Scrapy w Pythonie, napisany skrypt znajduje się w folderze crawler/spiders/links.py i content.py), a następnie przetworzone w taki sposób, aby otrzymać sam tekst (usunięcie tagów HTML'owych).

Słownik słów kluczowych został utworzony jako unia wszystkich słów występujących w wpisach. Przed dodaniem słowa do słownika jest ono poddane stemmingowi wykorzystując bibliotekę Natural Language Toolkit (NLTK), zatem w słowniku nie będą występować np. słowa argue, argued, argues, arguing tylko jedno słowo argu. Również są usuwane wszystkie słowa, które znajdują się w zbiorze stop_words wygenerowanym przez bibliotekę NLTK. Tak utworzony słownik w tym przypadku posiada 82308 słów.

Po utworzeniu macierzy wektorów cech term-by-document-matrix (kolumny odpowiadają dokumentom a wiersze danemu słowu) każdy jej wiersz został przemnożony przez inverse-document-frequency (IDF) danej następującym wzorem:

$$IDF(w) = \log \frac{N}{n_w},$$

gdzie N to liczba wszystkich dokumentów, a n_w to liczba dokumentów, w których występuje dane słowo *w*. Następnie każda kolumna została znormalizowana (jej długość wynosi jeden). Podana fraza zapytania jest zamieniana na odpowiedni wektor, który również jest normalizowany.

Po tych operacjach otrzymujemy następującą miarę prawdopodobieństwa:

$$|\mathbf{q}^T \mathbf{A}| = [|\cos\theta_1|, |\cos\theta_2|, \ldots, |\cos\theta_n|]$$

gdzie A to macierz wektorów cech. Po prawej stronie otrzymujemy wartość korelacji pomiędzy dokumentem n a danym zapytaniem.

Następny krok to usunięcie szumu wykorzystując SVD i low rank approximation. Naszą macierz A zapisujemy jako:

$$\mathbf{A} \simeq \mathbf{A}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T = [\mathbf{u}_1 | \dots | \mathbf{u}_k] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix} = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} {\mathbf{v_i}}^T$$

Kolumny tej macierzy również zostają poddane normalizacji, tak jak wcześniej.

W wyniku operacji SVD macierz A przestaje być macierzą rzadką i w tym przypadku posiada ok. 170 mln elementów niezerowych i zajmuje ok. 660 MB. Z tego względu nie jest ona załączona do sprawozdania. Dla porównania macierz wektorów cech zajmuje tylko 5 MB.

W przyszłości projekt zostanie rozszerzony, aby posiadał większą liczbę dokumentów i wykorzystywał Latent Dirichlet Allocation.

1. **Przykład dla zapytania ‘battle of kursk 1943’ bez SVD:**

Probabilities from sparse tbdm

(692, 0.05662224806610035)

http://ww2today.com/5th-july-1943-the-last-german-offensive-in-the-east-operation-citadel

(681, 0.0508476786288639)

http://ww2today.com/16th-july-1943-hitler-calls-off-the-operation-citadel-offensive

(689, 0.049321582509528186)

http://ww2today.com/8th-july-1943-a-german-view-of-a-panzer-attack-at-kursk

(691, 0.04796749824465721)

http://ww2today.com/6th-july-1943-a-soviet-artilleryman-blown-up-at-kursk

(709, 0.042699132816224346)

http://ww2today.com/18th-june-1943-a-prisoner-snatch-on-the-eastern-front

2. Porównanie wyników z SVD i bez SVD dla zapytania ‘invasion of normandy’.

Probabilities from sparse tbdm

(325, 0.12782606700336085)

http://ww2today.com/6-june-1944-0945-hitler-has-not-yet-been-told

(1812, 0.10889962308780628)

http://ww2today.com/26-june-1940-the-british-prepare-for-a-nazi-invasion

(384, 0.1065639935568889)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

(1799, 0.09518790915299132)

http://ww2today.com/10th-july-1940-churchill-considers-the-prospects-for-invasion

(483, 0.09102512237976076)

http://ww2today.com/29-january-1944-rommel-demands-stronger-defences-on-a-normandy-beach

Probabilities from svd dla k = 3

(51, 0.025907343798591234)

http://ww2today.com/22-february-1945-across-germany-in-the-special-custody-of-the-ss

(1863, 0.025903679523726283)

http://ww2today.com/16-may-1945-berliners-learn-to-accommodate-the-red-army

(1843, 0.025900168527231263)

http://ww2today.com/31-may-1945-for-millions-of-people-the-war-is-not-yet-over

(963, 0.025896984018544362)

http://ww2today.com/7th-october-1942-selected-to-live-in-treblinka

(852, 0.025892139332318452)

http://ww2today.com/26th-january-1943-desperate-fighting-in-stalingrad-but-no-surrender

Probabilities from svd dla k = 10

(348, 0.06309809792999044)

http://ww2today.com/5th-june-1944-2200-i-wish-to-god-it-were-safely-over

(1024, 0.06277093685793794)

http://ww2today.com/7th-august-1942-churchill-shakes-things-up-in-the-desert

(325, 0.0605067863489456)

http://ww2today.com/6-june-1944-0945-hitler-has-not-yet-been-told

(483, 0.05984519622211147)

http://ww2today.com/29-january-1944-rommel-demands-stronger-defences-on-a-normandy-beach

(727, 0.059566243701839924)

http://ww2today.com/31st-may-1943-churchill-argues-for-the-invasion-of-italy

Probabilities from svd dla k = 15

(344, 0.14214308425897587)

http://ww2today.com/6th-june-1944-0215-german-7th-army-still-undecided

(483, 0.13676281911710947)

http://ww2today.com/29-january-1944-rommel-demands-stronger-defences-on-a-normandy-beach

(334, 0.1359373213102295)

http://ww2today.com/6-june-1944-0700-utah-beach-assault-sustained

(348, 0.135765455202572)

http://ww2today.com/5th-june-1944-2200-i-wish-to-god-it-were-safely-over

(384, 0.13058022832357696)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

Probabilities from svd dla k = 20

(344, 0.14379311206337386)

http://ww2today.com/6th-june-1944-0215-german-7th-army-still-undecided

(334, 0.1395455618777756)

http://ww2today.com/6-june-1944-0700-utah-beach-assault-sustained

(348, 0.13888725429819226)

http://ww2today.com/5th-june-1944-2200-i-wish-to-god-it-were-safely-over

(332, 0.13480582359309473)

http://ww2today.com/6-june-1944-0725-tanks-land-in-advance-of-infantry-on-sword-beach

(384, 0.1338130443033989)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

Probabilities from svd dla k = 30

(344, 0.15393841368964653)

http://ww2today.com/6th-june-1944-0215-german-7th-army-still-undecided

(348, 0.1479902473576326)

http://ww2today.com/5th-june-1944-2200-i-wish-to-god-it-were-safely-over

(334, 0.14670252130604655)

http://ww2today.com/6-june-1944-0700-utah-beach-assault-sustained

(331, 0.14352074575201992)

http://ww2today.com/6-june-1944-0737-cruiser-hms-scylla-off-sword-beach

(338, 0.14003883842998266)

http://ww2today.com/6-june-1944-0558-daybreak-a-cold-grey-day-arrives

Probabilities from svd dla k = 50

(384, 0.15107338853687746)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

(348, 0.1494100994150511)

http://ww2today.com/5th-june-1944-2200-i-wish-to-god-it-were-safely-over

(338, 0.14815129729437942)

http://ww2today.com/6-june-1944-0558-daybreak-a-cold-grey-day-arrives

(483, 0.14622004167201821)

http://ww2today.com/29-january-1944-rommel-demands-stronger-defences-on-a-normandy-beach

(334, 0.14513713578645304)

http://ww2today.com/6-june-1944-0700-utah-beach-assault-sustained

Probabilities from svd dla k = 100

(384, 0.15849632586412285)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

(1812, 0.15311487125784695)

http://ww2today.com/26-june-1940-the-british-prepare-for-a-nazi-invasion

(338, 0.15158557298112726)

http://ww2today.com/6-june-1944-0558-daybreak-a-cold-grey-day-arrives

(325, 0.13803011679832872)

http://ww2today.com/6-june-1944-0945-hitler-has-not-yet-been-told

(1799, 0.13509082242151244)

http://ww2today.com/10th-july-1940-churchill-considers-the-prospects-for-invasion

Probabilities from svd dla k = 400

(325, 0.169730763292576)

http://ww2today.com/6-june-1944-0945-hitler-has-not-yet-been-told

(384, 0.15053055858463335)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

(1812, 0.14153138817541613)

http://ww2today.com/26-june-1940-the-british-prepare-for-a-nazi-invasion

(338, 0.1281587860068742)

http://ww2today.com/6-june-1944-0558-daybreak-a-cold-grey-day-arrives

(1913, 0.12502092179905117)

http://ww2today.com/german-view-of-the-invasion-of-norway

Probabilities from svd dla k = 1300

(325, 0.12569010358862923)

http://ww2today.com/6-june-1944-0945-hitler-has-not-yet-been-told

(1812, 0.10679033109077835)

http://ww2today.com/26-june-1940-the-british-prepare-for-a-nazi-invasion

(384, 0.10664967630206318)

http://ww2today.com/6-may-1944-us-low-level-photo-recon-surprises-germans

(483, 0.08948837642344923)

http://ww2today.com/29-january-1944-rommel-demands-stronger-defences-on-a-normandy-beach

(1799, 0.08726357522008842)

http://ww2today.com/10th-july-1940-churchill-considers-the-prospects-for-invasion

Według mnie najlepsze wyniki są dla k = 50 w przypadku tego pytania. Jednak wpisując inne pytania, np. ‘battle of kursk 1943’ wyniki są lepsze dla k = 1300.

Probabilities from svd dla k = 50 i zapytania ‘battle of kursk 1943’

(828, 0.0494792870885464)

http://ww2today.com/19th-february-1943-panzers-fail-in-second-assault-on-kasserine-pass

(955, 0.042431211682302045)

http://ww2today.com/15th-october-1942-the-unrelenting-battle-for-stalingrad-continues

(893, 0.041747056054869496)

http://ww2today.com/16th-december-1942-new-russians-tactics-delay-winter-storm

(1741, 0.040914744845485365)

http://ww2today.com/24th-august-1940-the-battleship-bismarck-is-commisssioned

(689, 0.040199472088907354)

http://ww2today.com/8th-july-1943-a-german-view-of-a-panzer-attack-at-kursk

Probabilities from svd dla k = 1300 i zapytania ‘battle of kursk 1943’

(692, 0.04476853866313571)

http://ww2today.com/5th-july-1943-the-last-german-offensive-in-the-east-operation-citadel

(689, 0.043648611626308946)

http://ww2today.com/8th-july-1943-a-german-view-of-a-panzer-attack-at-kursk

(681, 0.040427459837082826)

http://ww2today.com/16th-july-1943-hitler-calls-off-the-operation-citadel-offensive

(732, 0.03735681594270787)

http://ww2today.com/26th-may-1943-a-remote-outpost-of-raf-coastal-command

(675, 0.036266629900175575)

http://ww2today.com/22nd-july-1943-the-red-army-goes-on-to-the-offensive-after-kursk

3. Wpływ IDF na wyniki wyszukiwania dla zapytania 'battle of kursk' i k = 50.

Probabilities from sparse tbdm

(1822, 0.14085904245475275)

http://ww2today.com/churchill-the-battle-of-britain-is-about-to-begin

(1546, 0.1357813616483921)

http://ww2today.com/6th-march-1941-the-battle-of-the-atlantic-begins

(895, 0.12734290799340264)

http://ww2today.com/14th-december-1942-operation-winter-storm-pushes-on-towards-stalingrad

(1452, 0.12098347962395682)

http://ww2today.com/6th-june-1941-hitler-orders-all-soviet-commissars-to-be-shot

(1309, 0.11677484162422844)

http://ww2today.com/27th-october-1941-brest-bombed-again

Probabilities from svd

(1546, 0.08016282728568432)

http://ww2today.com/6th-march-1941-the-battle-of-the-atlantic-begins

(1822, 0.0763787729768406)

http://ww2today.com/churchill-the-battle-of-britain-is-about-to-begin

(1639, 0.0679491933944088)

http://ww2today.com/3rd-december-1940-fighter-command-still-active-by-day

(1477, 0.06755462688781698)

http://ww2today.com/14th-may-1941-second-eagle-squadron-formed

(895, 0.06657746529643914)

http://ww2today.com/14th-december-1942-operation-winter-storm-pushes-on-towards-stalingrad

Widzimy, że pojawiają się wyniki, które mają po prostu dużo jakiegoś słowa z zapytania, np. w dokumencie http://ww2today.com/churchill-the-battle-of-britain-is-about-to-begin słowo battle występuje aż 8 razy.

4. Poniżej kilka wyników wyszukiwania dla SVD i k = 50 i różnych zapytań.

Probabilities from svd dla zapytania ‘concentration camps’

(1494, 0.22400917893682942)

http://ww2today.com/27th-april-1941-himmler-visits-mauthausen

(779, 0.21175567233941714)

http://ww2today.com/9th-april-1943-welcome-to-gros-rosen-arbeit-macht-frei

(1591, 0.21079504005112543)

http://ww2today.com/20th-january-1941-himmler-visits-dachau

(504, 0.20590291523277768)

http://ww2today.com/9-january-1944-an-english-pow-at-work-in-auschwitz

(1045, 0.20389527492453385)

http://ww2today.com/17th-july-1942-auschwitz-the-sudden-death-of-yankel-meisel

Probabilities from svd dla zapytania ‘battle of britain

(1822, 0.12448237723001862)

http://ww2today.com/churchill-the-battle-of-britain-is-about-to-begin

(1746, 0.11435112170091297)

http://ww2today.com/20th-august-1940-never-in-the-field-of-human-conflict

(629, 0.10492408244648677)

http://ww2today.com/6th-september-1943-churchill-on-the-unity-of-the-english-speaking-peoples

(1594, 0.10109723883181866)

http://ww2today.com/17-january-1941-winston-churchill-harry-hopkins

(1373, 0.09965308010985438)

http://ww2today.com/24th-august-1941-churchill-the-power-of-the-english-speaking-peoples

Probabilities from svd dla zapytania ‘afrika korps’

(1222, 0.06004160606936937)

http://ww2today.com/21st-january-1942-rommels-surprise-attack-in-the-desert

(1029, 0.05875845733939698)

http://ww2today.com/2nd-august-1942-a-pause-in-the-desert-war

(1518, 0.05602464890595705)

http://ww2today.com/3rd-april-1941-rommels-first-success-in-the-desert

(1073, 0.054539235144091563)

http://ww2today.com/19th-june-1942-rommel-prepares-the-assault-on-tobruk

(1233, 0.052178485411272055)

http://ww2today.com/10th-january-1942-rommel-remains-confident-despite-retreat

Probabilities from svd dla zapytania ‘fall of france’

(1823, 0.09442135196461897)

http://ww2today.com/18th-june-1940-de-gaulle-declares-that-the-fight-goes-on

(1825, 0.0922038215675959)

http://ww2today.com/16th-june-1940-civilians-continue-to-flee-the-war

(232, 0.08877851654901603)

http://ww2today.com/25-august-1944-paris-broken-paris-martyred-but-paris-liberated

(855, 0.08635066256702487)

http://ww2today.com/23rd-january-1943-the-battle-of-marseille

(1784, 0.08466895337605428)

http://ww2today.com/24th-july-1940-french-liner-meknes-torpedoed

Probabilities from svd dla zapytania ‘war on the eastern front’

(1089, 0.09573568876634299)

http://ww2today.com/3rd-june-1942-the-eastern-front-settles-down-to-trench-warfare

(1254, 0.09490848172801346)

http://ww2today.com/20th-december-1941-hitler-appeals-for-warm-clothing-for-eastern-front-troops

(1278, 0.09351857923729906)

http://ww2today.com/27th-november-1941-the-russian-winter-arrives-on-the-eastern-front

(1235, 0.0907843025850285)

http://ww2today.com/8th-january-1942-a-german-counter-attack-on-the-eastern-front

(809, 0.08998263910038326)

http://ww2today.com/10th-march-1943-fresh-wehrmacht-troops-encounter-the-rasputitsa