

Sprawozdanie - Latent Semantic Indexing

Arkadiusz Kraus

21 maja 2019

1 Zadanie

Zadanie polegało na zastosowaniu metody Latent Semantic Indexing (LSI), aby w zbiorze tekstów móc wyszukiwać te związane z podaną frazą.

2 Zbiór danych

Zbiór danych jaki zastosowałem pochodzi ze strony <https://www.kaggle.com/snapcrack/all-the-news>. Jest to zbiór artykułów z amerykańskich gazet z lat 2016-2017 o różnej tematyce. Jest złożony z ponad 140 tys. tekstów.

3 Zastosowane przekształcenia

Na początku przetwarzamy wstępnie przetwarzamy zbiór artykułów, aby później móc skuteczniej w nim wyszukiwać:

3.1 Zbiór użytych słów

Dla każdego artykułu bierzemy wszystkie użyte słowa i tworzymy wektor słów dla wszystkich tekstów. Dla kolejnych podzbiorów artykułów posiada on następujące wielkości.

liczba artykułów	rozmiar wektora
1000	44 tys.
10000	140 tys.
50000	230 tys.
wszystkie(142572)	400 tys.

3.2 Wstępne przetworzenie

Liczba słów uzyskana w poprzednim podpunkcie jest bardzo duża, aby uczynić algorytm wydajniejszym staramy się zredukować ten wektor. We wstępnym przetworzeniu pomijamy wielkość liter oraz znaki interpunkcyjne.

3.3 Stemming

Wiele słów może występować w różnych formach np. take i taking. Nas jednak interesuje sam fakt czynności, a nie koniecznie np. czas w jakim była wykonywana. Innym przypadkiem jest też liczba mnoga rzeczownika, gdzie nadal interesuje nas sam rzeczownik. Dlatego stosujemy algorytm stemmingu (dokładniej Porter Stemmer), doprowadzamy słowa do wspólnego rdzenia i usuwamy duplikaty. W połączeniu z poprzednim krokiem pozwala to znacząco zredukować rozmiar wektora:

liczba artykułów	rozmiar wektora
1000	24 tys.
10000	70 tys.
50000	127 tys.
wszystkie(142572)	234tys.

3.4 Stopwords

Stopwords to zbiór słów takich jak przyimki itp, które występują bardzo często we wszelkich tekstach. Nie mają one większego znaczenia dla treści dlatego również można je pominąć. Nie redukuje to wielkości wektora znacznie (o około 150 słów), jednak zwiększa jakość wyszukiwania.

3.5 Tworzenie macierzy rzadkiej

Chcemy teraz utworzyć macierz, która jako wiersze będzie przyjmować kolejne artykuły, a jako kolumny kolejne słowa ze zbioru. Wartością w komórce $a_{i,j}$ macierzy będzie liczba wystąpień j-tego słowa w i-tym artykule. Zauważmy, że będzie to macierz rzadka ponieważ artykuły mają ok. 4000 tys słów, a wektor słów jest znacznie większy. Dodatkowo jeśli artykuł jest na jakiś temat to słowa często się w nim powtarzają. Aby przyspieszyć obliczenia (i w ogóle je umożliwić) stosujemy więc przechowywanie w postaci macierzy rzadkiej (dokładniej CSR - Compressed Sparse Row).

3.6 IDF

W celu zwiększenia jakości wyszukiwania stosujemy Inverse Document Frequency (IDF). Pozwala on nam zmniejszyć znaczenie słów, które występują w wielu artykułach i zwiększyć tych, które rzadko.

3.7 Normalizacja

Normalizacja pozwala uniezależnić korelację od długości tekstu. Wykonujemy ją od razu, aby potem przyspieszyć wyszukiwanie, aby nie trzeba było obliczać normy za każdym razem.

3.8 Odszumianie

W celu redukcji szumów stosujemy algorytm SVD i wybieramy k najciekawszych wartości własnych. Co ciekawe zwiększa to liczbę niezerowych elementów macierzy z ok 1-2 mln do 87 mln dla 10 tys. artykułów.

4 Obliczenia

Problem ten jest bardzo wymagający obliczeniowo oraz pamięciowo. Przechowanie macierzy wielkości $liczba\ artykułów * liczba\ słów$ dla większej ilości artykułów skutkuje skończeniem pamięci. W pliku *logs* znajdują się czasy wstępnego przetworzenia dla 50000 i wszystkich artykułów. Wyszukiwanie po przetworzeniu i odczytaniu wcześniejszej macierzy jest również zależne od ilości artykułów i dla większej ilości jeszcze trochę mu brakuje względem czasów osiągniętych przez Google.

liczba artykułów	rozmiar wektora	rozmiar wektora po przetworzeniu	czas przetworzenia	czas wyszukania
1000	44 tys	24 tys	natychmiast	natychmiast
10000	140 tys	70 tys	1 min	1 s
50000	230 tys	127 tys	35 min	15 sek
wszystkie(142572)	400 tys	234 tys	2,5h	1 min

Niestety dla 50000 oraz dla wszystkich artykułów stworzenie macierzy po dekompozycji SVD okazało się zbyt czasochłonne. Przy próbie utworzenia normalnej macierzy po dekompozycji dla np. 200 wartości własnych otrzymywałem brak pamięci, natomiast dla macierzy rzadkich obliczenia trwały w nieskończoność (obliczenia na nich są czasochłonnymi operacjami). Dlatego dla tych wielkości wyszukiwarka używa macierzy bez odszumienia.

5 Wyniki

Aplikacja jest podzielona na dwie części - jedna odpowiadająca za indeksowanie i druga wyszukująca treści dla zadanej frazy. Obie części działają poprawnie. Poniżej znajduje się przykład:

weather rain south

Search

Hurricane Matthew Toll Climbs to at Least 17 as North Carolina Suffers Record-Breaking Flooding - The New York Times

Correlation: 0.08182207146267585

Hurricane Matthew was downgraded to a cyclone early Sunday morning as it hit North Carolina and Virginia with a weakened but still powerful punch. Dispatches from our reporters on the ground a live storm tracker map and answers to reader questions will be updated below. ■ The storm's death toll in the United States has climbed to at least 17. Gov. Pat McCrory of North Carolina said on Sunday morning that his state's toll had risen to eight. Officials in Georgia confirmed three deaths on Saturday. At least six fatalities in Florida have been attributed to the storm. Nearly 900 people have died in Haiti, according to a Reuters report based on information from civil protection and local officials. ■ Bands of heavy rain are leading to flooding in parts of eastern North Carolina, according to the National Hurricane Center, which may result in flooding and flash flooding elsewhere in the region. Forecasters warned that areas along the Neuse River in Goldsboro, N. C. could experience flooding worse than the devastating inundation that followed Hurricane Floyd in 1999. The river was projected to hit 30.9 feet around 2 a. m. on Monday, surpassing the record of 28.9 feet caused by Floyd. ■ In Georgia, the storm created a record surge at Tybee Island, near the state's border with South Carolina. The surge reached 12.5 feet, according to the Chatham County Emergency Management Agency, which exceeded the previous high of 12.2 feet, set during Hurricane David in 1979. ■ To cover the storm and its aftermath, The New York Times has journalists deployed along the its path. Follow our correspondents on Twitter. The storm's assault on North Carolina extended into Sunday, and the governor said at least seven people had been killed in the state. "I wouldn't assume that there aren't people clinging for life right now in houses that are underwater that we have yet to reach, especially in areas," Gov. Pat McCrory told reporters in Raleigh, the state capital. "That's what my major concern is." Heavy rain was still pounding parts of the state as the governor spoke, and Mr. McCrory said floodwaters could rush through North Carolina for days. The fallout, he said, "is going to be a prolonged event." "This is still an extremely dangerous situation," Mr. McCrory said. "And I cannot stress it more especially in the areas of Rocky Mount, Kinston, Greenville, Goldsboro and other eastern towns, cities. We're going to have major issues with rivers and flooding." Mr. McCrory said 58 boat crews had rescued nearly 900 people by Sunday morning. More than 4,000 people were in shelters on Sunday, Mr. McCrory said, and about 760,000 homes and businesses were without electricity. Parts of Interstate 40 and Interstate 95 were shut down. — ALAN BLINDER The storm lashed South Carolina's Lowcountry with rain and sent the sea surging upward. It turned roads into rivers and ripped trees from the inundated ground before making landfall north of Charleston. More than 430,000 people around the state had been left without power as of late Saturday. Gov. Nikki Haley of South Carolina said on Sunday that she had lifted evacuation orders for four counties, including Charleston. But residents in four other counties — Beaufort, Georgetown, Horry and Jasper — were still urged to stay away. "We had a lot of rain, for many hours," said Capt. Bob Bromage, the public information officer for the Sheriff's Department in Beaufort County, southwest of Charleston. "We had flooding, we may have had some surge that we're not aware of yet, we have downed trees, we have reports of downed trees on houses already." — JESS BIDGOOD, in Charleston The American Red Cross has published a list of safety tips for homeowners returning to property damaged in hurricanes. Once cleanup begins, taking photographs of damaged or destroyed property before it is discarded provides a record for use in making insurance claims, says Ann Carms, who writes the Your Money Adviser consumer finance column for The Times. Gov. Rick Scott of Florida said on Saturday that the state had suffered "unbelievable amounts of beach erosion," as well as damaged roads, flooding and felled trees. But Mr. Scott, who flew along the coastline to survey storm damage, said the storm's consequences could have been far worse. "The first thing we can say is we are all blessed that Matthew stayed off our coast," he said. "I worried the whole time that even though the track was off our coast, that it would turn in and have a direct hit at some point." State and local officials said they were trying to expedite recovery. "We're going to ride each other hard," said Lenny Curry, the mayor of Jacksonville. "We're going to ride our utilities, we're going to ride all our workers to get this community back together as quickly and as safely as possible." Still, Mayor Charlie Latham of Jacksonville Beach asked for patience. *"Matthew saved lives in Florida, and he has fought with them. You've taken our lives, but lost to our people's homes and provided safe service." Mr. T. Auburn said. AT A N. H. INT'D in Orlando in 14:01 the*

Tropical Storm Hermine Leaves Trail of Power Failures - The New York Times

Correlation: 0.08009828004945116

The remnants of what had been Hurricane Hermine swept up and along the Eastern Seaboard on Saturday, disrupting the holiday weekend in much of the coastal South while preparing to bedevil the Northeast well into the week. The storm, which made its landfall early Friday near St. Marks, Fla. as a Category 1 hurricane, led to hundreds of thousands of power failures and flooded roadways. In Florida, dealt its first direct strike by a hurricane in nearly 11 years, the Tallahassee area was particularly hard hit, and officials said it could be almost a week before electricity was fully restored. "We still have a lot of work to do following the storm," Gov. Rick Scott of Florida said at a news conference on Saturday in Tallahassee, the state capital. "We'll continue to spend the coming days assessing the damage and responding to the needs of our Florida families." Mr. Scott said there was "significant damage" in the state, including "a lot of downed power lines." Residents and officials described destroyed businesses, boats set adrift, crumbled sea walls and battered homes. The National Hurricane Center said the storm, whose winds reached as high as 80 miles per hour after it spent days swirling through the Gulf of Mexico, remained powerful on Saturday, even after it crossed the Florida Panhandle, Georgia and the Carolinas. The storm became a cyclone on Saturday morning, but forecasters warned that it still "could be near hurricane intensity" for part of the week. And so the consequences of the storm's path stretched from Florida, which reported the first fatality attributed to it, to New York, where the city said beaches would be closed Sunday, to Cape Cod, Mass. which was under a tropical storm watch. The hurricane center said Saturday evening that the storm would "meander slowly offshore of the coast for the next couple of days" before churning past New Jersey and New York. Officials said storm surges would be "accompanied by large and dangerous waves," and they warned of the possibility of "inundation" from Virginia to Connecticut. Gov. Chris Christie of New Jersey declared emergencies for Atlantic, Cape May and Ocean Counties, and Gov. Andrew M. Cuomo of New York said the state's emergency operations center would be activated on Sunday. On Saturday, the National Guard was on alert in Virginia, where floodwaters prompted road closings in the Hampton Roads region, and visitors to North Carolina's Outer Banks, where strong rip currents were reported, were advised to adjust their travel schedules. "I'm relieved that everyone took this storm seriously, paid attention to the weather forecast and heeded the warnings of state and local officials," said Gov. Pat McCrory of North Carolina, whose state shut down major bridges in the Outer Banks because of winds associated with the storm. Officials said that the death of one person in a traffic accident in eastern North Carolina was probably connected to the storm. Although people in many communities said the storm would soon become only a minor memory, parts of Florida were entering what could be a protracted recovery, and there were some concerns that the absence of could prove dangerous. In Leon County, which includes Tallahassee, more than half of homes were without electricity on Saturday, and the American Red Cross opened a comfort shelter, a place "for residents to cool off, hydrate and receive information from area nonprofit agencies." The authorities planned to distribute bottled water throughout the county, where high temperatures were predicted to be in the upper 80s or lower 90s for the next several days. During his appearance on Saturday at Florida's emergency operations center, Mr. Scott made a dispiriting roll call of counties and the extent of their power failures. In Wakulla County, just south of Tallahassee, about 72 percent of customers were without electricity. "This is the No. 1 issue people are telling me about," Mr. Scott said before leaving for another afternoon of damage assessments.

Hurricane Matthew's Toll Rises Flooding Strands 1,500 in North Carolina - The New York Times

Correlation: 0.07677973382307797

Rysunek 2: Wyniki wyszukiwania dla hasła "weather rain south" część druga

Dla większej ilości artukulów gdzie nie było możliwe przeprowadzenie SVD zdarzają się jednak wyniki takie jak np. ten:

Want to be the next Mark Zuckerberg? His CEO sister says don't do these 3 things

Correlation: 0.27399432445989885

" 'Fans of "Shark Tank" know there are some. ...'

Sometimes They Wink At You - Breitbart

Correlation: 0.1842246519195048

I can't imagine what Sony's marketing department was trying to tell us about fans of the new feminist Ghostbusters.

This is what happens when you slice a laptop in half with water

Correlation: 0.17145244151302674

"Or, if you're looking for some fun, slicing through everyday things like shoes or a laptop." 'Business Insider talked to the creator of a new YouTube channel called "... This person didn't want to be identified since they are using the tool at work — albeit — but they did walk us through what it's like to turn water into something that can slice a baseball cleanly in half.' 'It may look like a just a stream of ...'

Watch: Vin Scully Says Goodbye After 67-Year Career - Breitbart

Correlation: 0.1665685942552541

📷 📖 pic.twitter. Legendary baseball announcer Vin Scully signed off for the final time Sunday after the Los Angeles fell to the San Francisco Giants . He said goodbye after a and reassured baseball fans everywhere that life will go on without him. Transcript as follows: "You know, friends, so many people have wished me congratulations on a career in baseball and they wished me a wonderful retirement with my family and now, all I can do is tell you what I wish for you. " 'May God give you.. For every storm, a rainbow, For every tear, a smile, For every care, a promise, And a blessing in each trial. For every problem life sends, A faithful friend to share, For every sigh, a sweet song, And an answer for each prayer.' You and I have been friends for a long time, but I know in my heart

Rysunek 3: Przykład wyniku z szumami

W tym przypadku dopiero któryś wynik jest taki jak powinien, a na początku jest szum. To samo zapytanie przy mniejszej ilości artykułów i z usuniętym szumem wygląda następująco:

Tim Tebow Homers in First At-Bat (and Finishes 1 for 6) - The New York Times

Correlation: 0.08417560679864274

Tim Tebow stepped to the plate Wednesday for his first as a professional baseball player and took a swing at the first pitch he saw. Naturally, he hit a home run. Tebow is a polarizing figure in the sports world. As a Heisman Trophy winner while at the University of Florida and, briefly, as an N. F. L. quarterback, he attracted both adulation and scorn for his overt Christian faith and his habit of Tebowing, or bowing in prayer after successes. On the field, some saw a quarterback with anemic passing statistics while others saw a player who always seemed to find a way to win. He faced considerable skepticism when, with his football days seemingly over, he announced plans to try for a baseball career at age 29. He had not played the game since high school. Still, scouts from 28 teams attended an open workout, and the Mets saw enough to sign Tebow to a minor league contract and send him to the Florida Instructional League. Game No. 1 came Wednesday against the St. Louis Cardinals' team in front of about 250 fans in Port St. Lucie, Fla. Tebow batted second and played left field. His first ended in seconds with the homer to . The pitcher was John Kilichowski, drafted out of Vanderbilt by the Cardinals in the 11th round in June. He posted a 2.70 E. R. A. in Class A with the Peoria Chiefs and the State College Spikes this past season, giving up four home runs in 11 games. Tebow's teammates rushed out to greet him at home plate after the bash. At that point, his career stats looked like this: batting average 1.000, average 1.000, slugging average 4.000. And let's make up a stat. Ratio of pitches seen to home runs delivered: 1:1. Barry Bonds? Babe Ruth? Tim Tebow was surpassing them all, and making the skeptics look pretty foolish. In his second Tebow grounded into a double play, dropping his batting average to a merely superhuman .500. Then he was 1 for 3, and 1 for 4, and he finished the day 1 for 6.

One Season Ends and Another Begins: Baseball Playoff Matchups Are Set - The New York Times

Correlation: 0.07783777397365577

The baseball gods spend six months twisting the sport into a knotty lump. No one knows quite how to untangle it, and then the final week zips by. All those possibilities, resolved just like that. The season rushes to the end. "Rushes to the end," Baltimore Orioles Manager Buck Showalter said, musing in his office at Yankee Stadium before this weekend's critical series. "That sounds like a great name for a book." Now we can close the volume on the 2016 regular season. The Orioles survived to live another day, earning a date in Toronto on Tuesday for the American League game against the Blue Jays. The Mets will host the National League game the next night, against the San Francisco Giants at Citi Field. The Orioles and the Blue Jays both won on Sunday to finish with identical records. The Blue Jays won the teams' season series to earn the right to host the game, which was also how the Mets did it. The Mets and the Giants both finished but the Mets won the season series. The A. L. winner will face the Texas Rangers, who went in a division series opener on Thursday. The Boston Red Sox will visit Cleveland that day to start their series with the Indians. The N. L. series begin Friday in Chicago and Washington, with the Cubs () facing the winner and the Nationals hosting the Los Angeles Dodgers. Before we get there, though, the appetizers will be delicious. The Orioles lost the A. L. Championship Series in 2014. The Blue Jays lost it last year. Neither of the teams, rivals in the A. L. East, has won a pennant in decades, and both are trying to do it with brute force. The Orioles led the majors in homers this season, with 253, and the Blue Jays had 221 — the next highest total in the playoff field. The N. L. game features a dream pitchers' duel: the Mets' Noah Syndergaard against the Giants' Madison Bumgarner. In Syndergaard's last postseason appearance — Game 3 of last fall's World Series against Kansas City — he announced his presence with a

Rysunek 4: Przykład wyniku bez szumu

6 Wnioski