Kolokacje, trochę gramatyki i reprezentowanie słów

Paweł Rychlikowski

Instytut Informatyki UWr

20 listopada 2018

Wnioskowanie statystyczne

- Twórz hipotezę H_0 że wystąpienie wspólne dwóch słów jest zupełnie przypadkowe (nie ma związku między tymi słowami)
- Oblicz prawdopodobieństwo, że (zakładając H_0) częstości wystąpień pary są takie jak obserwujemy i następnie odrzuć H_0, jeżeli p wyjdzie małe (0.05, 0.01).
- Dla nas ten przypadek bez związku, to zdarzenia niezależne (zatem $P(w_1 w_2) = P(w_1)P(w_1)$)

Test t-Studenta

- Test t (sttudenta): H_0 oznacza, że pobraliśmy próbkę z czegoś o rozkładzie o średniej μ
- liczymy wartość: $t = \frac{\overline{x}-\mu}{\sqrt{\frac{s^2}{N}}}$. μ to średnia z próbki, s^2 – wariancja.
  Dla $\alpha = 0.005$ odrzucamy H_0 (czyli: jest kolokacja) jeżeli $t > 2.576$
- Myślimy o korpusie jako o długim ciągu par, a dla każdej pary mamy 0/1 mówiące, czy to nasza para.
- Odchylenie standardowe (rozkład Bernouliego) to jest $s^2 = N * p * (1 - p)$, czyli w przybliżeniu Np.
- Daje to wzór:

$$t \approx \frac{p(w_1 w_2) - p(w_1)p(w_2)}{\sqrt{p(w_1 w_2)}}$$

Pointwise Mutual Information. Przypomnienie

- Inną opcją jest PMI
- Sortujemy wg

$$\log\left(\frac{P(w_1 w_2)}{P(w_1)P(w_2)}\right)$$

Positive Pointwise Mutual Information jest równe $\max(0, PMI(w_1, w_2))$

- czyli jak coś jest rzadziej niż „przypadkowo” to dajemy 0

PMI. Problem z rzadkimi słowami

- Dla rzadkich słów wartość PPMI jest trochę zbyt duża.
- Stosowanym rozwiązaniem jest zmiana sposobu zliczania kontekstów:
  - zamiast $P(c)$ stosujemy $P_\alpha(c) = \frac{\text{cnt}(w)^\alpha}{\sum_w \text{cnt}(w)^\alpha}$
  - Dobrze działa $\alpha = 0.75$
  - Zwiększamy w ten sposób p-stwo rzadkich zdarzeń (a zatem zmniejszamy PMI).

Inne wzorki

- Łatwo zauważyć, że funkcji rozsnących ze względu na 1 argument i malejących ze względu na dwa pozostałe jest dużo.
- Można popatrzeć na pracę:

  O.Kolesnikova, Survey of Word Co-occurrence Measures for Collocation Detection

Uwaga

We wzorkach często porównujemy liczby wystąpień, rzeczywistą oraz szacowaną przy zależeniu nieazależności wystąpnień.

Przykładowy wzorek

Poisson-Stirling Approximation

$$\text{PSM} = f(xy) \cdot (\log f(xy) - \log \hat{f}(xy) - 1)$$

gdzie $f(xy)$ to liczba wystąpień bigramu xy, a $\hat{f}(xy) = \frac{f(x)f(y)}{N}$ (czyli szacowana liczba wystąpień)

W wielu sytuacjach można osiągnąć lepsze rezultaty, zakładając niezerowe wykorzystanie wiedzy lingwistcznej.

Na przykład w zadaniu kolokacji (uwzględnienie gramatyki może pomóc „statystyce")

A teraz trochę koniecznej lingwistyki

Czego uczyli nas w szkole?

- Każdy wyraz jest jakąś częścią mowy.
- Główne części mowy to rzeczownik, czasownik, przymiotnik, przysłówek.
- Istnieją też inne części mowy, takie jak przyimek, spójnik, zaimek, partykuła.
- Podział na części mowy zawdzięczamy Dionizusowi Thraxowi z Aleksandrii (ok 100pne). Wyodrębnił on 8 wyżej wymienionych części mowy (bez partykuły, ale za to z rodzajnikiem).

Przykłady

Standardowe części mowy

1. Rzeczownik: krowa, koń, sytuacja, uczucie
2. Czasownik: być, mieć, robić
3. Przymiotnik: ładny, piękny, najurodziwszy
4. Przysłówek: ładnie, pięknie, najurodziwiej, bardzo
5. Imiesłów (jak widać różne warianty): umierając, umierający, umarłszy, umarły, zabijany (!umierany)

Pozostałe części mowy

1. Przyimek: do, poprzez, od, wokół, niczym
2. Zaimek: on, jego, mój, tak, taki, ile, gdzie
3. Spójnik: i, oraz, lecz, lub, że
4. Liczebnik: dwa, trzy, czwarty
5. Rodzajnik: a, the, der, die, das, eine, les
6. Inne dziwne (wykrzykniki, partykuły, kubliki,

Wybrane problemy z częściami mowy

Problem z zaimkiem

Stefan nie rozumiał Judyty. **Jej** idee nie w ogóle do **niego** nie przemawiały.

Konieczne może być określenie, do czego odnosi się dany zaimek.

Problem z przyimkiem

The doctor examined the man with a sthetoscope.

The doctor examined the man with a broken leg.

examined with vs **man with**

Do czego przyda się znajomość POS

Dygresja

Często będziemy używać skrótu POS (part of speech).

- Po angielsku czytacz powinien je znać: DIScount vs disCOUNT, CONtent vs conTENT
- Język jako ciąg tagów POS (niedokładne przybliżenie, ale czasem wystarczające)
- Tekst wzbogacony o POS tagi jest bardziej jednoznaczny,

  *Spraw (**rozkaźnik?** rzeczownik?) się dobrze w ministerstwie Spraw(rozkaźnik? **rzeczownik?**) Zagranicznych.*
  *Jasiu je (**czasownik?** zaimek?) słonecznik i śmieci (**rzeczownik?** czasownik?)*

Otwarte czy zamknięte

- Najbardziej zgrubny podział to klasy otwarte oraz zamknięte.
- Otwarte: pojawiają się ciągle nowe słowa, bo na przykład odkrywamy nową substancję, lub wynajdujemy urządzenie.
- Zamknięte: raczej nie wynajdziemy dodatkowego spójnika lub przyimka.

Otwarte części mowy

Rzeczowniki, przymiotniki, czasowniki, przysłówki, przysłówki, imiesłowy. Przykład generowania:

guglować, wyguglować, guglowanie, guglowy, guglując, guglujący, wyguglowawszy, guglowo...

Kolokacje z filtrem gramatycznym

Filtrujemy gramatycznie najczlepsze (najczęstsze?) bigramy. Jak?

- Przeglądamy bigramy przechodzące test statystyczny
- Dla każdego słowa patrzymy w słowniku na wszystkie możliwe części mowy.
- Pozostawiamy tylko:
  - rzeczownik przymiotnik
  - przymiotnik rzeczownik
  - rzeczownik czasownik
  - czasownik rzeczownik
  - rzeczownik rzeczownik

O znaczeniu słów

Przeprowadzimy eksperyment, w którym spróbujemy wydedukować znaczenie słów

Znaczenie słowa w zależności od kontekstu. Rzeczownik

Wolę karmić **X?** na wolności , są wtedy prawdziwe .
Typowym żywicielem jest **X?** Funambulus tristriatus .
Z mniejszych ssaków występuje lis , łasica , **X?** , zając .
Stąd nie nadają się do podglądania życia seksualnego **X?** .

Znaczenie słowa w zależności od kontekstu. Rzeczownik

Na jesienną słotę najlepsza jest sycąca **X?** z mnóstwem witamin i minerałów .
Czasem **X?** wychodzi zbyt rzadka lub brak jej ostatecznego smaku .
Uwielbiam **X?** dyniową , ale marynowanej dyni nie zdzierżę !
Spożywać bezpośrednio lub z mlekiem , jogurtem , kefirem , **X?** .

Znaczenie słowa w zależności od kontekstu. Przymiotnik

Ciebie **X?** chłopców otacza czereda ,
Dorodne te śliweczki - chętnie bym zjadła , róże masz **X?**.
Kumczo , **X?** lilie ... wogóle ogród piękny ....
Mam do sprzedania **X?** szczeniaczki rasy sznaucer miniatura .

Znaczenie słowa w zależności od kontekstu. Czasownik

Delikatny dotyk **X?** skórę , poprawia cyrkulację krwi i limfy .
Pielęgnacja przeciwzmarszczkowa , która ujędrnia , **X?** , odżywia skórę dojrzałą .
Zabieg poprzedzony jest peelingiem , który złuszcza naskórek i **X?** stopy .
Dopełniła się ofiara , a morze **X?** , bo tak chciał Pan .

Rozwiązanie

Brakujące słowa to:

1. wiewiórka
2. zupa
3. prześliczny
4. wygładzić

Zanurzenia wektorowe

Cel

Chcemy patrząc na konteksty stworzyć wektorową reprezentację słowa.

Pożądane właściwości

1) Wektory mają stały wymiar
2) Patrząc na dwa wektory jesteśmy w stanie powiedzieć, czy słowa są podobne
3) Podobieństwo słów to podobieństwo kontekstów w których występują
4) Idealnie: wektory mają niezbyt duży wymiar

Rzadkie zanurzenia wektorowe

Uwaga

Rezygnujemy z wymogu niewielkiej wymiarowości!

Reprezentację wektorową wyznaczamy następująco:

a) Uznajemy, że kontekstem słowa jest inne słowo (albo inny lemat)

b) Decydujemy się na określony słownik, jego wielkość będzie wymiarem wektora

c) Dla wektora $\mathbf{x}$ będącego reprezentacją słowa łasica mamy:
   - wartości $\mathbf{x}_{\text{pozycja(kuna)}} > 0$, $\mathbf{x}_{\text{pozycja(futerko)}} > 0$
   - wartość $\mathbf{x}_{\text{pozycja(albebra)}} = 0$, $\mathbf{x}_{\text{pozycja(krab)}} = 0$

d) Wartości niezerowe to może być na przykład PPMI

e) Wektory można znormalizować, czyli podzielić przez długość.

Podobieństwo wektorów

- Podobne wektory powinny mieć niezerowe wartości na mniej więcej tych samych pozycjach.
- Sprawdza to całkiem dobrze iloczyn skalarny:

$$x \cdot y = \sum_{i=1}^{N} x_i y_i$$

- Jak wektory są znormalizowane, to iloczyn skalarny jest cosinusem.

Częstym sposobem wyznaczania podobieństwa jest podobieństwo cosinusowe (czyli iloczyn skalarny znormalizowanych wektorów)

Gęste wektory

- Wolelibyśmy mieć wymiar rzędu 100-1000 a nie milion.
- Rzadkie wektory są nieco wolniejsze od gęstych (niezerowych wartości może być sporo)

Ważniejszy problem

Osie nie mają ze sobą związku: niezerowa wartość na pozycji hotel w jednym wektorze nie dopasuje się do wartości na pozycji motel

Uwaga

Można próbować łączyć osie, ale nie da się w ten sposób łatwo uwzględnić hierarchii synonimów
szerszeń, modliszka, tygrys, pantera (są drapieżnikami, ale tworzą też dwie naturalne pary)

Word2vec: wariant Skip Grams with Negative Sampling

- Kontekstem jest nieodległe słowo (stąd skip-grams)
  - (około 5 wyrazów maksymalnie)
- Chcemy nauczyć klasyfikator odpowiadać na pytanie:

  Czy słowo t pasuje do kontekstu c?

- Zakładamy, że klasyfikator ma dostęp do podobieństwa wektorów dla słowa i dla kontekstu, którym jest iloczyn skalarny odpowiednich wektorów.
- Chcemy obliczać: $P(+|t, c)$ jako funkcję podobieństwa.

Word2vec (2)

- Sposobem na zamianę wartości na prawdopodobieństwo jest funkcja sigmoidalna, czyli:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- Czyli:

$$P(+|t, c) = \frac{1}{1 + e^{-t \cdot c}}$$

- Oczywiście:

$$P(-|t, c) = 1 - P(+|t, c) = \frac{e^{-t \cdot c}}{1 + e^{-t \cdot c}}$$

Word2vec (3)

- Mamy pozytywne przykłady słów w kontekście (to nasz korpus).
- Negatywne po prostu losujemy (dwa losowe słowa raczej do siebie nie pasują)
  - Losowanie negatywnych ze zmodyfikowanymi prawdopodobieństami, wg $P_{\alpha=0.75}(c)$, bo inaczej trochę zbyt często the.

Cel

Tak dobrać wektory, żeby prawdopodobieństwo (suma logarytmów) pozytywnych przykładów było jak największe, a negatywnych jak najmniejsze

Word2Vec (4)

- Musimy zmaksymalizować następującą funkcję:

$$L(\theta) = \sum_{(t,c)\in+} \log P(+|t,c) + \sum_{(t,c)\in-} \log P(-|t,c)$$

  gdzie θ jest zbiorem parametrów (czyli wektorami słów i kontekstów dla całego korpusu)
- W praktyce sprowadza się to do zbliżania dobrych par i oddalania złych (podczas przeglądania całego korpusu).

Reklama

Więcej na sieciach neuronowych (że to zmniejszanie jest najpierw szybsze, potem wolniejsze, że to jest algorytm Stochastic Gradient Descent, i wiele innych szczegółów z tym związanych.

Demonstracja wektorów (1)

- Wektory nauczone na fragmencie Wikipedii (ok. 500 MB)
- Dla form bazowych (po prostu losowana forma bazowa dla słów wieloznacznych)

Uwaga

To trochę za mały korpus, używa się kilkukrotnie większych.

Wyniki word2vec (zlematyzowana Wikipedyjka)

```
WORD krowa
    0.747756484098 świnia
    0.741241809754 trzoda
    0.737357741294 bydło
    0.735480104802 nierogacizna
    0.734951379794 świnić
    0.733063170364 owca
    0.716620248075 pasący
    0.701427074831 klacz
    0.700166953566 indyk
    0.700021406894 osioł
    0.695569407282 jagnię
    0.694016069978 wielbłąd
    0.69152733642 juczny
    0.691095780167 hodować
    0.68366141303 pojenie
    0.681089444495 wierzchowiec
    0.680008274911 koza
    0.679638822666 koń
```

Wyniki word2vec (zlematyzowana Wikipedyjka)

```
WORD miłość
    0.78878611847 namiętność
    0.776644681332 samotność
    0.768324552769 tęsknota
    0.765030976931 radość
    0.757115660981 uczucie
    0.748768628658 pragnienie
    0.748452868438 grzeszny
    0.748112227824 szaleństwo
    0.741791104905 uczuć
    0.73309407139 platoniczny
    0.729588293624 marzyciel
    0.727683475742 rozterka
    0.724713659762 siostrzyczka
    0.722602825599 zakochany
    0.7222541249 smutek
    0.721232759442 zauroczenie
    0.71700863921 dziewczyna
    0.71638479031 szczęście
```

Przykładowe zastosowanie

Zadanie

Upraszczanie tekstu – czyli zamieniamy tekst na prostszy wariant, zachowując (częściowo) sens i poprawność gramatyczną.

Jak to zrobić i po co?

Przykładowe zastosowania

1. Wyszukiwarka fraz w korpusie (tłumaczymy na polski uproszczony i znajdujemy frazy bliskoznaczne, bo o tym samym tłumaczeniu)
2. Model językowy na dużo mniejszej liczbie słów (zobaczymy jak wygląda uproszczenie zdania: babuleńka miała dwa rogate koziołki).

Przykładowe zastosowanie (2)

Rozwiązanie

Zamieniamy słowa na słowa o tej samej charakterystyce gramatycznej i bliskich zanurzeniach, znajdujące się wśród Top K najpopularniejszych słów.

babuleńka miała dwa rogate koziołki
grupka miała dwa kamienne ptaszki
babuleńka — grupka halinka panienka ciotka latarnia
rogate — kamienne przedziwne maleńkie dzikie przepiękne
koziołki — ptaszki orły zające słonie węże

judyta podarowała wczoraj stefanowi czekoladki
ewelina przekazała wczoraj jerzemu kanapki
judyta — ewelina marta weronika ola natalia
podarowała — przekazała kupiła przywiozła wręczyła zapłaciła
stefanowi — jerzemu józefowi andrzejowi tadeuszowi pawłowi
czekoladki — kanapki sałatki czekolady obraczki butelki

Z lematami byłoby fajniej

polowanie na nosorożce włochate powinno być dozwolone w niektórych rezerwatach
polowanie na słonie bose powinno być dozwolone w niektórych lasach
nosorożce — słonie ptaki zające lwy węże
włochate — bose białe rude wilgotne maleńkie
rezerwatach — lasach terenach obszarach rejonach ogrodach

ubój nosorożców włochatych powinien być dozwolony w niektórych rezerwatach
przemyt komarów nagich powinien być zakazany w niektórych lasach
ubój — przemyt handel proceder kilogram transport
nosorożców — komarów ptaków trupów owadów lisów
włochatych — nagich wilgotnych siwych mokrych cienkich
dozwolony — zakazany uregulowany ograniczony przewidziany dopuszczony
rezerwatach — lasach terenach obszarach rejonach ogrodach