Zastosowanie metod statystycznych do ekstrakcji słów kluczowych w kontekście projektu LT4eL

Łukasz Degórski

LT4eL

Language Technology for e-Learning

Wykorzystanie narzędzi językowych oraz technik sieci semantycznej do wspomagania wyszukiwania materiałów dydaktycznych.

- Automatyczna ekstrakcja słów kluczowych
- Automatyczna ekstrakcja kontekstów definiujących
- Podłączenie do ontologii, a dzięki temu: wyszukiwanie materiałów zbliżonych merytorycznie mimo niewystępowania szukanego klucza, oraz wyszukiwanie materiałów w innych językach niż w zapytaniu

Słowa kluczowe – punkt wyjścia

Kolekcja dokumentów – tekstów do badania, być może mocno zróżnicowanej wielkości.

Każdy dokument powinien zostać opatrzony słowami kluczowymi, oddającymi jego treść.

Słowo nie jest precyzyjnym określeniem, mogą to być ciągi słów (bigramy, trigramy,...) typu *analiza numeryczna algorytmu*. Dla ustalenia uwagi będziemy mówić jednak o słowach.

Bag of words assumption

Przy wielu zastosowaniach metod statystycznych zakłada się, że badana kolekcja dokumentów jest takim “workiem” i analizuje wyłącznie liczbę wystąpień słowa w całym worku.

To się łatwo liczy i modeluje (jednoparametrowy rozkład - Poissona lub dwumianowy), ale próby np. modelowania liczby dokumentów w kolekcji, w jakich dane słowo wystąpi (document frequency) wykazały, że taki prosty model się nie sprawdza.

Powstały bardziej dokładne i skomplikowane modele – Two Poisson, Negative Binomial, K-mixture, ale...

Nie działa – i całe szczęście

...efektem ubocznym tych badań było spostrzeżenie, że dla niektórych słów uproszczony model nie działa bardziej, a dla niektórych mniej.

Różnica między liczbą dokumentów faktyczną a modelowaną jest dużo większa dla rozkładających się nierównomiernie między dokumentami słów związanych z treścią (*content words*), niż dla słów funkcyjnych (*function words*).

Słowa mocno związane z treścią są doskonałymi kandydatami na **słowa kluczowe** (mówiące o czym jest tekst).

Dlaczego tak jest?

Im mniej słowo zachowuje się przewidywalnie (czyli da się modelować rozkładem Poissona), tym większe podejrzenie, że mają na to wpływ jakieś *ukryte zmienne*, take jak tematyka tekstu.

Tymczasem rozkład Poissona oraz rozkład dwumianowy zakładają niezależność wystąpień.

Dla “interesujących” słów warunek ten nie jest spełniony. Jedno wystąpienie zwiększa, a nie zmniejsza prawdopodobieństwo kolejnego: jeśli w tekście wystąpią słowa *rozkład poissona*, to tekst zapewne dotyczy statystyki, a więc powinniśmy bardziej się spodziewać jego kolejnych wystąpień niż w tekście ogólnym.

Konkrety konkretne

Mając kolekcję ***D*** dokumentów definiujemy dla danego słowa *w*:

$\boldsymbol{df_w}$: liczba dokumentów, w których to słowo wystąpiło (*document frequency*)

IDF: $-log_2(\boldsymbol{df_w} / \boldsymbol{D})$ (*inverse document frequency*)

Dla wybranego dokumentu, $\boldsymbol{df_w} / \boldsymbol{D}$ jest prawdopodobieństwem, że w dokumencie wystąpi przynajmniej raz słowo *w*.

Konkrety modelowane

$\pi(\theta,k)$: Prawdopodobieństwo wg rozkładu Poissona z parametrem θ, że w dokumencie słowo w wystąpi k razy.

θ ma być średnią liczba wystąpień w na dokument, obliczamy ją dzieląc liczbę wystąpień w całej kolekcji $\boldsymbol{cf}_w$ przez $\boldsymbol{D}$.

Stąd modelowane IDF równe jest:

$$\boldsymbol{IDF}_{\boldsymbol{\pi}} = -log_2(1-\pi(\theta,0))$$

czyli, na mocy powyższego oraz definicji rozkładu Poissona:

$$\boldsymbol{IDF}_{\boldsymbol{\pi}} = -log_2(1-e^{-\boldsymbol{cf}_w/\boldsymbol{D}})$$

Konkrety - podsumowanie

Znając D, df_w, cf_w możemy więc obliczyć IDF oraz IDF_π i je porównywać.

Analizujemy tzw. *residual IDF*, czyli różnicę

$$IDF - IDF_\pi$$

Słowa, przy których dostaniemy dużą różnicę (model się wyraźnie pomyli), uznajemy za słowa kluczowe.

Można w podobny sposób oprzeć się na innych niż IDF miarach, np. wariancji, entropii, adaptacji.

Problemy badawcze

- Jak wyraźne zróżnicowanie wielkości dokumentów wpływa na wyniki – czyli czy wystąpienie wystąpieniu równe?
- Które miary najlepiej zastosować do porównania?
- Czy warto analizować też błędy dokładniejszych modeli?
- Jaki powinien być próg “kluczowości”?

Podróż do wnętrza dokumentu

Dotychczas przeszliśmy od badania liczby wystąpień w całej kolekcji (bag of words) do badania liczby wystąpień w poszczególnych dokumentach.

Kolejnym krokiem jest zbadanie rozmieszczenia kandydatów na słowa kluczowe wewnątrz poszczególnych dokumentów – miejsc wystąpień oraz tendencji do ich skupiania.

Nawet często występujące słowa funkcyjne nie rozkładają się równomiernie wewnątrz dokumentu; słowa związane z treścią tym bardziej nie.

Modelowanie - odstępy

Budujemy osobny model dla konkretnego słowa w w całej kolekcji. Dla każdego dokumentu mamy ciąg $w_1 ... w_k$ odstępów między wystąpieniami tego słowa.

Zakładamy, że słowo występuje z bazowym prawdopodobieństwem $1/\alpha$, ale gdy już wystąpi, prawdopodobieństwo kolejnego wystąpienia w niewielkiej odległości zwiększa się do $1/\beta$.

Stąd, współczynnik powtarzalności (*rate of re-occurence*) jest modelowany przez sumę dwóch rozkładów wykładniczych: z dużą średnią (dla pierwszego wystąpienia) oraz z mniejszą (dla powtórzonych).

Model

$$\phi_1(w_j) = \alpha e^{-\alpha w_j}$$

$$\phi_2(w_j) = \beta e^{-\beta w_j}$$

Niech p będzie prawdopodobieństwem załapania się do pierwszego rozkładu. Stąd:

$$\phi(w_j) = p\phi_1 + (1 - p)\phi_2$$

Warunki brzegowe

Przy pierwszym wystąpieniu drugi wykładniczy składnik znika:

$$\phi(w_1) = \phi_1(w_1) = \alpha e^{-\alpha w_1}$$

Przy ostatnim wystąpieniu: *cenzurowanie* (technika z badań klinicznych) – jeśli obserwacja (kolejne wystąpienie słowa) nie zaszła, zakładamy, że zaszłaby kiedyś, gdyby dokument się nie skończył. Czyli – ostatni badany odstęp jest równy odległości ostatniego wystąpienia od końca dokumentu.

Podobnie, jeśli słowo nie występuje w ogóle, zakładamy, że wystąpiłoby, gdyby dokument miał nieskończoną długość – jedynym badanym "odstępem" jest długość dokumentu.

Podejście bayesowskie

Zamiast zakładać, że parametry dystrybucji są stałe, a dane się zmieniają, przyjmujemy odwrotnie. Rozpoczynając od założeń nie niosących żadnej informacji, w kolejnych krokach otrzymujemy coraz lepsze przybliżenia parametrów.

Parametry naszego modelu: $\boldsymbol{p}$, α, β

Założenia wstępne:

$\boldsymbol{p} \sim \text{Unifom}(0,1)$
$\alpha \sim \text{Uniform}(0,1)$
$\beta = \alpha + \omega$, gdzie $\omega \sim \text{Uniform}(0,1)$ - żeby zawsze $\beta > \alpha$

Wyniki, interpetacja

Po wyliczeniu parametrów (Gibbs sampling) dla każdego słowa otrzymujemy wartości $\boldsymbol{p}$, α, β.

Intuicyjna interpretacja:
α małe, β małe: często występujące, typowe słowo funkcyjne
α małe, β duże: dość częste, ale rozrzucone słowo funkcyjne
α duże, β małe: słowo związane z treścią, skupiające wystąpienia
α duże, β duże: rzadkie i rozrzucone słowo funkcyjne

Praktycznie:
Należy analizować iloraz α/β i od pewnego przyjętego progu uważać słowo za kluczowe.

Problemy badawcze

- Jaki dobrać próg?
- Czy stosować obie techniki naraz, a jeśli tak – z jakimi wagami brać pod uwagę ich wyniki)?

Nie tylko statystyka

Potencjalne pozastatystyczne modyfikatory oceny "kluczowości":

- **Formatowanie** miejsca wystąpienia: dodatkowy bonus za wystąpienie w nagłówku, pogrubieniu, ...

- Informacja **morfosyntaktyczna**: zdecydowany bonus dla fraz rzeczownikowych pasujących do wybranych wzorców

- Ewentualnie **heurystyki**, np. **T**rzy**L**itero**S**króty

- ...

Bibliografia

- Church, K., Gale, W. *Poisson Mixtures*
- Church, K., Gale, W. *Inverse Document Frequency (IDF): A Measure of Deviations form Poisson*
- Sarkar, A., Garthwaite, P.H., De Roeck, A. *A Bayesian mixture model for term re-occurrence and burstiness*
- Katz, S.M., *Distribution of content words and phrases in text and language modelling*