

Co sieć złożona może powiedzieć o tekście?

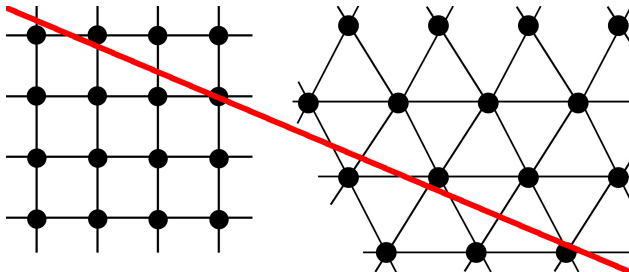
Tomasz Stanisław

Zakład Teorii Systemów Złożonych, IFJ PAN, Kraków

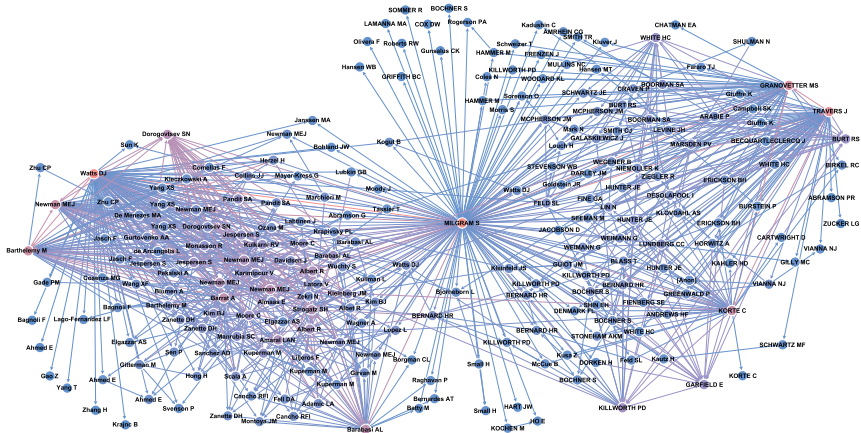
7 października 2019

Sieci złożone

Sieć złożona - "graf o nietrywialnych właściwościach topologicznych"



Sieci złożone



Sieci złożone w badaniach języka

Przykłady sieci lingwistycznych:

- sieci reprezentujące podobieństwo znaczeń słów
- sieci oparte o relacje gramatyczne pomiędzy słowami
- sieci przedstawiające współwystępowanie słów w tekście

Przykłady zastosowań:

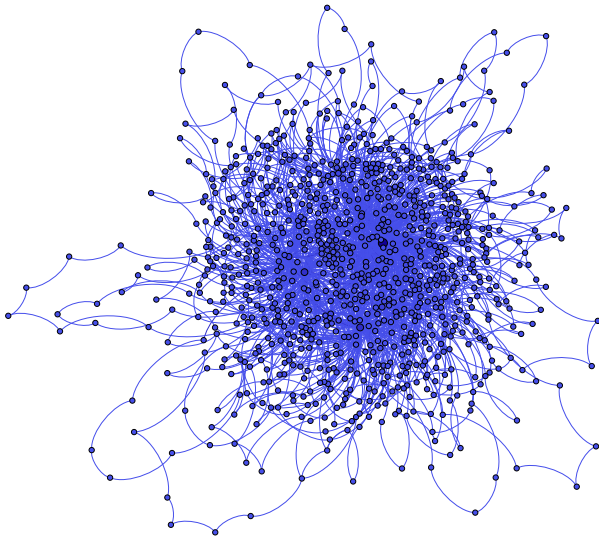
- ujednoznacznianie
- tłumaczenie maszynowe
- wybór słów kluczowych dokumentu

Sieci sąsiedztwa słów

Nieważona sieć sąsiedztwa słów:

- każde unikalne słowo w tekście $\Leftrightarrow$ węzeł sieci
- para słów pojawiających się w tekście obok siebie $\Leftrightarrow$ krawędź sieci

Sieci sąsiedztwa słów



Sieci sąsiedztwa słów

Wykorzystanie sieci sąsiedztwa słów w badaniu języka:

- ocena zgodności tekstu z przyjętą konwencją za pomocą parametrów sieci;
- ocena jakości tłumaczeń automatycznych;
- rozpoznawanie autorstwa tekstu - np. poprzez badanie dynamiki sieci lub zliczanie "motywów sieciowych".

Sieci sąsiedztwa słów

Pomysł: przebadąć najprostsze, standardowe wielkości charakteryzujące sieci sąsiedztwa słów i sprawdzić, czy można za ich pomocą na przykład zidentyfikować autora tekstu.

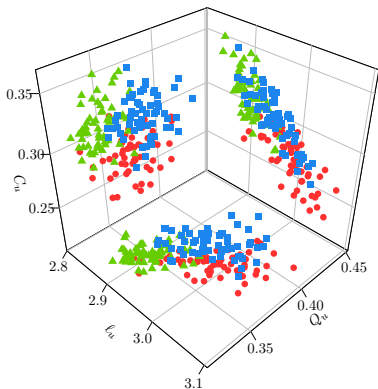
Przygotowanie tekstów

- Zbiór tekstów: 48 książek w języku angielskim i 48 w polskim: w każdym z języków po 8 autorów, po 6 książek na każdego autora.
- W każdym tekście usunięte zostają: przedmowa, przypisy, komentarz do wydania, itp..
- Znaki interpunkcyjne: kropka, pytajnik, wykrzyknik, wielokropek, przecinek, średnik, pauza, dwukropek oraz lewy i prawy nawias zostają zamienione na symbole, które w analizie są traktowane tak, jakby były słowami. Pozostałe znaki występujące w tekście zostają usunięte.

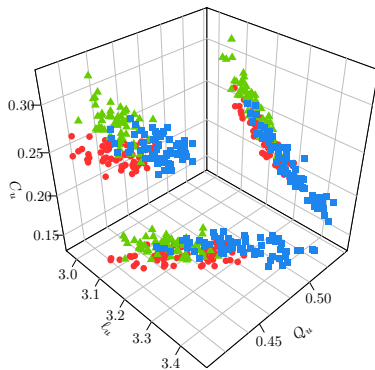
Tworzenie sieci z tekstu

- Z każdego tekstu tworzymy dwie sieci sąsiedztwa słów - jedną nieważoną i jedną ważoną. Wszystkie sieci są **nieskierowane** i tylko takie są przedmiotem zainteresowania.
- Badamy charakterystyki sieci.

Analiza skupień



G. Eliot, J. Austen, J. Conrad



H. Sienkiewicz, J. Kraszewski, S. Żeromski

Stopień wężła v (w sieciach ważonych także *siła* wężła v):

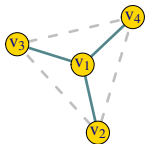
- w sieciach nieważonych jest to liczba krawędzi wychodzących z wężła, oznaczenie: $\deg(v)$;
- w sieciach ważonych jest to suma wag krawędzi wychodzących z wężła, oznaczenie: $\text{str}(v)$;

Lokalne charakterystyki sieci

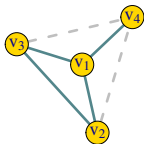
Współczynnik gronowania (klastrowania, klasteryzacji) wężła v :

- w sieciach nieważonych jest to prawdopodobieństwo, że dwa dowolnie wybrane węzły połączone krawędzią z węzłem v są połączone krawędzią także między sobą; jeżeli m_v oznacza liczbę wszystkich krawędzi, których oba końce są węzłami połączonymi krawędzią z węzłem v , to współczynnik gronowania wężła v wyraża się wzorem:

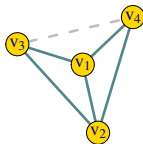
$$C_u(v) = \frac{2m_v}{\deg(v) \cdot (\deg(v) - 1)};$$



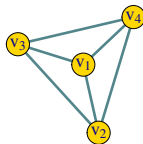
$$C_u(v_1) = 0$$



$$C_u(v_1) = 1/3$$



$$C_u(v_1) = 2/3$$



$$C_u(v_1) = 1$$

Lokalne charakterystyki sieci

- uogólnień współczynnika gronowania na sieci ważne jest kilka, jedno z nich wyrażone jest wzorem:

$$C_w(v) = \frac{1}{\text{str}(v) \cdot (\deg(v) - 1)} \sum_{u,t \in S(v)} \frac{w_{vu} + w_{vt}}{2} a_{vu} a_{ut} a_{tv},$$

gdzie $S(v)$ jest zbiorem węzłów połączonych krawędzią z węzłem v , w_{uv} oznacza wagę krawędzi między węzłami u i v (gdy nie ma takiej krawędzi, to $w_{uv} = 0$), zaś

$$a_{uv} = \begin{cases} 1, & \text{jeżeli w sieci istnieje krawędź między węzłami } u \text{ i } v; \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Wedle tej definicji współczynnik gronowania węzła v opisuje wkład do siły węzła v od tych par węzłów, które razem z nim tworzą zamknięty trójkąt.

Lokalne charakterystyki sieci

Długość $d(u, v)$ ścieżki z wężła u do wężła v to:

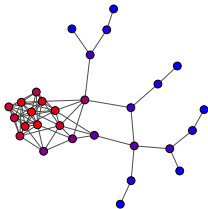
- w sieci nieważonej: liczba krawędzi na tej ścieżce;
- w sieci ważonej: suma odwrotności wag na tej ścieżce.

Średnia długość $\ell(v)$ najkrótszej ścieżki wężła v : średnia arytmetyczna z długości najkrótszych ścieżek między wężłem v a wszystkimi innymi wężłami sieci:

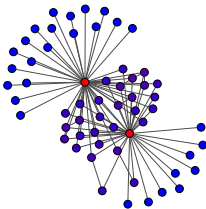
$$\ell(v) = \frac{1}{N-1} \sum_{u \in V \setminus \{v\}} d(v, u) \quad (N - \text{liczba wężłów sieci}).$$

Globalne charakterystyki sieci

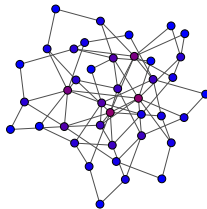
Współczynnik asortatywności r - współczynnik korelacji pomiędzy stopniami (w sieci ważonej: siłami) węzłów połączonych krawędzią.



$$r = 0.74$$



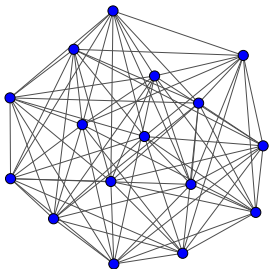
$$r = -0.82$$



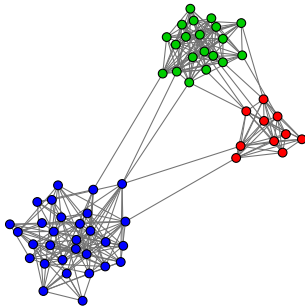
$$r \approx 0$$

Globalne charakterystyki sieci

Modularność Q - stopień w jakim sieć może być podzielona na rozłączne zbiory węzłów, takie, że wewnątrz tych zbiorów znajduje się wiele krawędzi, natomiast krawędzi łączących węzły z różnych zbiorów jest niewiele.



$$Q = 0.07$$



$$Q = 0.58$$

Globalne charakterystyki sieci

Istnieją też globalne charakterystyki sieci oparte o uśrednienie charakterystyk lokalnych po całej sieci:

- globalny współczynnik gronowania - średnia arytmetyczna ze współczynników gronowania wszystkich węzłów;
- globalna średnia długość najkrótszej ścieżki - średnia arytmetyczna z długości najkrótszych ścieżek między wszystkimi parami węzłów.

Normalizacja charakterystyk

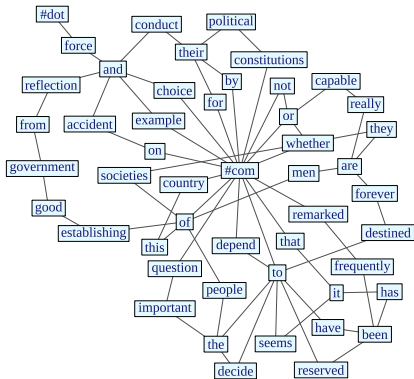
Aby skompensować efekty będące konsekwencją zróżnicowanej długości tekstów oraz ewentualnie te wynikające wprost z rozkładu częstości słów, wszystkie badane wielkości były odpowiednio znormalizowane.

- Z oryginalnego tekstu tworzymy sieć, i obliczamy zadaną wielkość, na przykład globalny współczynnik gronowania, C .
- Oryginalny tekst poddajemy randomizacji - losowo przedstawiamy w nim kolejność słów. Z tak powstałego tekstu tworzymy sieć i obliczamy interesującą nas wielkość.
- Randomizację powtarzamy pewną ilość razy (np. 50). Z każdego powtórzenia uzyskujemy jedną wartość badanej charakterystyki.

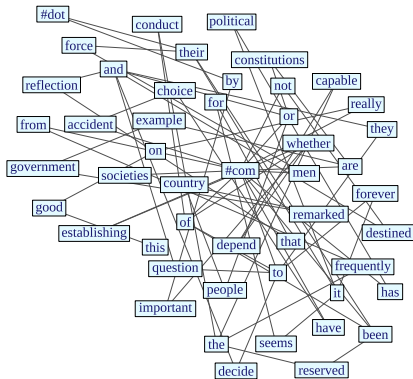
Normalizacja charakterystyk

- Uzyskane wartości uśredniamy, otrzymując, średnią, typową wartość charakterystyki dla sieci powstałej z tekstu randomizowanego; w rozważanym przykładzie będzie to średni randomizowany współczynnik gronowania, C_{rand} .
- Znormalizowaną charakterystykę określamy jako iloraz jej wartości dla tekstu oryginalnego i średniej wartości dla tekstu randomizowanego; dla współczynnika gronowania będzie to:
$$\tilde{C} = C / C_{rand}.$$
- Wyjątek - siła wężła! Znormalizowaną siłę wężła określamy jako $\widetilde{\text{str}}(v) = \text{str}(v) / \sum_{u \in V} \text{str}(u)$, gdzie V jest zbiorem wężłów sieci.

Normalizacja charakterystyk



It has been frequently remarked, that it seems to have been reserved to the people of this country, by their conduct and example, to decide the important question, whether societies of men are really capable or not, of establishing good government from reflection and choice, or whether they are forever destined to depend, for their political constitutions, on accident and force.

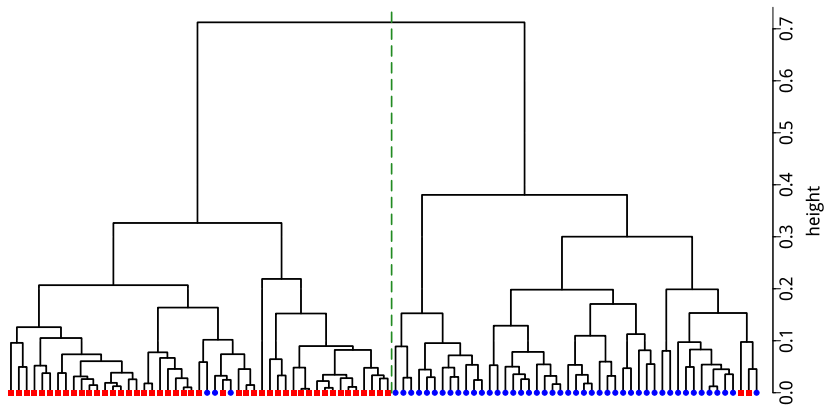


this good on really , societies , to question and they to decide and destined not , men force their . by important whether of to of conduct the frequently from , are example government remarked country people or constitutions are , been reserved the capable depend whether establishing , of been their It political has , have choice accident , or and reflection that for seems it forever to

Po obliczeniu odpowiednich wielkości, z jednego tekstu uzyskujemy wektor liczb (globalne parametry sieci, parametry lokalne dla wybranego zbioru słów). Do wyciągnięcia informacji ze zbioru takich wektorów stosujemy standardowe metody uczenia maszynowego.

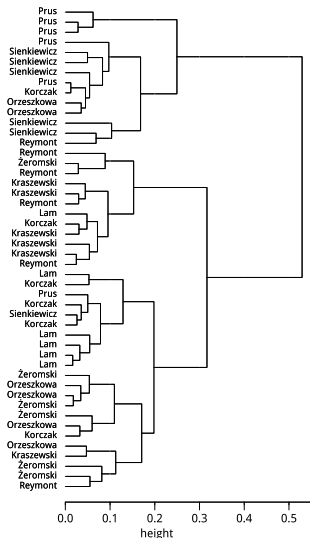
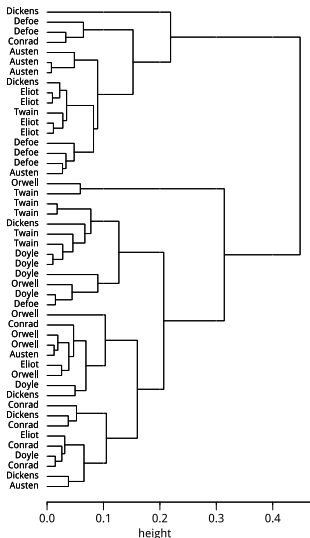
Analiza skupień

Grupowanie hierarchiczne wszystkich tekstów (EN, PL) w przestrzeni globalnych charakterystyk sieci nieważonych (wsp. gronowania, wsp. asortatywności, średnia długość najkrótszej ścieżki, modularność); czerwony - EN, niebieski - PL.



Analiza skupień

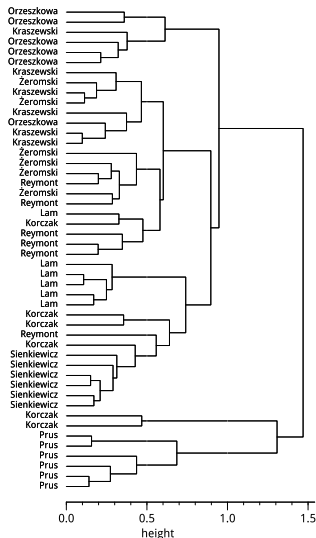
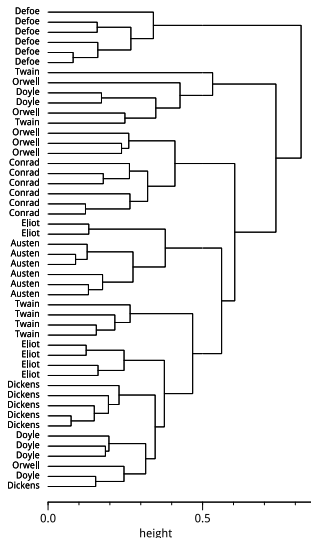
Grupowanie hierarchiczne tekstów (lewy rysunek - EN, prawy - PL) w przestrzeni globalnych (znormalizowanych) charakterystyk sieci nieważonych.



Badanie charakterystyk lokalnych: w każdej sieci wybieramy pewien zbiór węzłów - na przykład tych odpowiadających n najczęściej występującym słowom, obliczamy dla nich odpowiednie parametry lokalne i otrzymane zbiory liczb traktujemy jako dane wejściowe do algorytmów grupowania lub klasyfikacji.

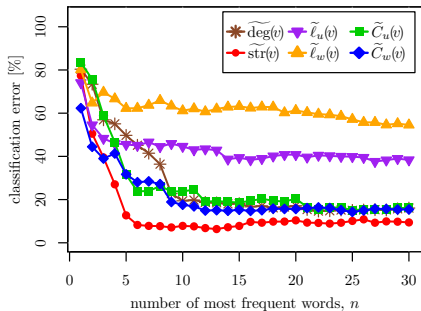
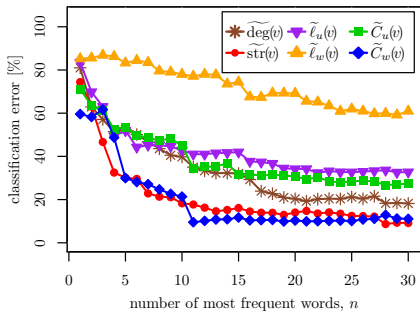
Analiza skupień

Grupowanie hierarchiczne tekstów (lewy rysunek - EN, prawy - PL) w przestrzeni (znormalizowanych) ważonych współczynników klastrowania 12 najczęstszych słów.

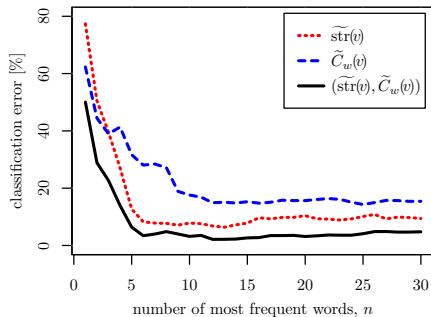
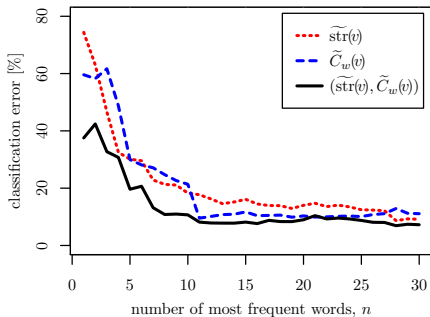


Klasyfikujemy teksty ze względu na autora, w przestrzeni lokalnych charakterystyk sieciowych. Z 6 tekstów każdego autora wybieramy losowo 4 do treningu algorytmu (zespół drzew klasyfikacyjnych), a pozostałe 2 wędrują do zbioru testowego. Wybór zbioru trenującego i trening klasyfikatora powtarzamy wielokrotnie. Z uśrednionych po wszystkich powtórzeniach wyników klasyfikacji w zbiorze testowym otrzymujemy prawdopodobieństwo poprawnej klasyfikacji.

Klasyfikacja



Klasyfikacja



Klasyfikacja

Wyniki klasyfikacji (w zbiorze testowym) tekstów (lewa tabela - EN, prawa - PL), w przestrzeni (znormalizowanych) ważonych współczynników klastrowania 12 najczęstszych słów. Liczba w i -tym wierszu i j -tej kolumnie oznacza prawdopodobieństwo zaklasyfikowania tekstu i -tego autora jako tekstu j -tego autora.

	Au	Co	De	Di	Do	El	Or	Tw
Au	.96	.00	.00	.01	.01	.02	.00	.00
Co	.00	.92	.00	.00	.00	.04	.04	.00
De	.00	.00	1.0	.00	.00	.00	.00	.00
Di	.01	.00	.00	.88	.04	.00	.00	.07
Do	.01	.00	.00	.05	.93	.01	.00	.00
El	.02	.02	.00	.03	.01	.87	.02	.03
Or	.00	.13	.00	.01	.01	.07	.78	.00
Tw	.00	.01	.00	.00	.01	.02	.07	.89

Au - Austen, Co - Conrad, De - Defoe, Di - Dickens, Do - Doyle, El - Eliot, Or - Orwell, Tw - Twain.

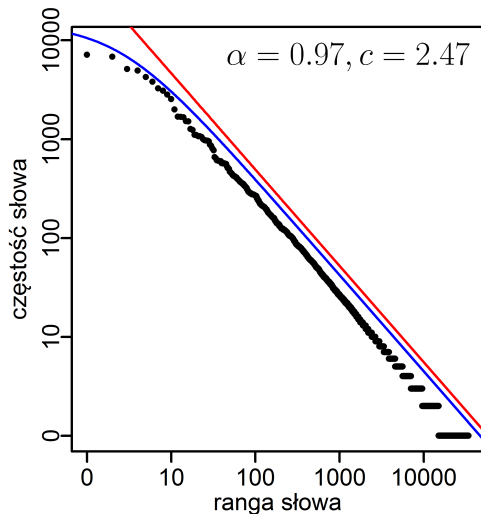
	Ko	Kr	La	Or	Pr	Re	Si	Że
Ko	.73	.00	.00	.00	.00	.06	.03	.18
Kr	.00	.80	.00	.12	.00	.00	.05	.03
La	.00	.00	1.0	.00	.00	.00	.00	.00
Or	.00	.15	.00	.83	.00	.00	.01	.01
Pr	.00	.00	.00	.00	1.0	.00	.00	.00
Re	.04	.00	.00	.00	.00	.77	.02	.17
Si	.00	.05	.00	.00	.04	.01	.90	.00
Że	.01	.15	.00	.00	.00	.03	.02	.79

Ko - Korczak, Kr - Kraszewski, La - Lam, Or - Orzeszkowa, Pr - Prus, Re - Reymont, Si - Sienkiewicz, Że - Żeromski.

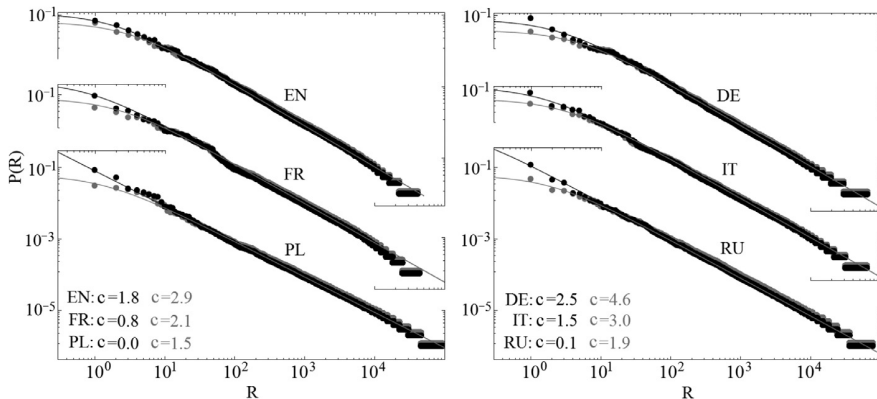
- Poprawność klasyfikacji ogółem: rzędu 90%.
- Pominięcie interpunkcji (nawet przy zachowaniu liczby atrybutów) prowadzi do pogorszenia wyników.
- Klasyfikacja z użyciem tylko kropki i przecinka daje skuteczność rzędu 50%.

Prawo Zipfa

- Zipf: $f \propto \frac{1}{R^\alpha}$
($\alpha \approx 1$)
- Zipf-Mandelbrot:
 $f \propto \frac{1}{(R+c)^\alpha}$
($\alpha \approx 1$)



Prawo Zipfa - z interpunkcją



- Przydatność sieciowej reprezentacji języka.
- Możliwość łączenia z innymi metodami.
- Rola interpunkcji.



T. Stanisław, J. Kwapien, S. Drożdż,

Linguistic data mining with complex networks: A stylometric-oriented approach,

Information Sciences 482 (2019) 301–320.



A. Kulig, J. Kwapien, T. Stanisław, S. Drożdż,

In narrative texts punctuation marks obey the same statistics as words,

Information Sciences 375 (2017) 98–113.

Angielski:

- Arthur Conan Doyle
- Charles Dickens
- Daniel Defoe
- George Eliot
- George Orwell
- Jane Austen
- Joseph Conrad
- Mark Twain

Polski:

- Bolesław Prus
- Eliza Orzeszkowa
- Henryk Sienkiewicz
- Jan Lam
- Janusz Korczak
- Józef Ignacy Kraszewski
- Stefan Żeromski
- Władysław Reymont

Dodatki - wsp. asortatywności

In unweighted networks, the assortativity coefficient can be defined as the Pearson correlation coefficient between the degrees of nodes that are connected by an edge. Let (u, v) denote the ordered pair of vertices that are connected by an edge. Since edges are undirected and the pair (u, v) is ordered, two such pairs can be assigned to each edge in the network. For each pair one can calculate the degrees of vertices u and v , and form a pair $(\deg(u), \deg(v))$. The set of all pairs $(\deg(u), \deg(v))$ for all edges can be treated as the set of values of a certain two-dimensional random variable (X, Y) . With such notation, the assortativity coefficient r_u is expressed by the Pearson correlation coefficient of variables X and Y :

$$r_u = \text{corr}(X, Y).$$

The generalization of the above formula to weighted networks is done in this paper by replacing the degrees of vertices by their strengths, and calculating the weighted correlation coefficient instead of normal one. Let (X, Y) be a two-dimensional random variable whose values are pairs $(x, y) = (\text{str}(u), \text{str}(v))$ for all pairs of vertices (u, v) connected by an edge. Let w be a function that to each pair $(x, y) = (\text{str}(u), \text{str}(v))$ assigns the weight of an edge connecting u and v . Then the weighted assortativity coefficient r_w can be written as:

$$r_w = \text{wcorr}(X, Y; w),$$

where $\text{wcorr}(X, Y; w)$ denotes the Pearson weighted correlation coefficient of variables X and Y with the weighing function w .

Since the assortativity coefficient is expressed by the correlation coefficient, it has values between -1 and 1. Networks with positive r are assortative, while networks with negative r are disassortative.

Dodatki - modularność

Modularity is a global characteristic of a network, measuring the extent to which the set of network's vertices can be divided into disjunctive subsets which maximize the density of edges within them, and minimize the number of edges connecting one with another. In word-adjacency networks, it can be interpreted as divisibility of the set of words appearing in the text into groups of words that frequently co-occur with each other.

Consider an unweighted network, with the set of vertices V . By *partition of the network* we understand the division of V into disjoint subsets (called modules, clusters or communities). Let a_{uv} denote the adjacency matrix element. Let c_v denote the module to which the vertex v is assigned by the given partition. The *modularity of a partition* is defined as:

$$q_u = \frac{1}{2m} \sum_{u,v \in V} \left(\left[a_{uv} - \frac{\deg(u) \deg(v)}{2m} \right] \delta(c_u, c_v) \right),$$

where m is the number of the edges of the network, $\deg(u)$, $\deg(v)$ are degrees of vertices u and v , and function $\delta(c_u, c_v)$ has value 1 if $c_u = c_v$ and 0 otherwise.

Modularity of a partition has value between -1 and 1, and indicates whether the density of edges within the given groups is higher or lower than it would be if edges were distributed at random. The random network that serves as a reference in this definition can be constructed using the so-called *configuration model*.

The *modularity of a network* Q_u is the maximum value from modularities q_u of all possible partitions. Determining the network's modularity precisely is computationally intractable, hence a number of heuristic algorithms have been proposed. In this paper modularity is calculated using the Louvain method.

The generalization of modularity onto weighted networks can be done by replacing the quantities appearing in Eq. 40 by their weighted counterparts. If w_{uv} denotes the weight of the edge connecting vertices u and v , M is the sum of all edge weights, and $\text{str}(u)$, $\text{str}(v)$ are the strengths of vertices u , v , then the modularity of a given partition is equal to:

$$q_w = \frac{1}{2M} \sum_{u,v \in V} \left(\left[w_{uv} - \frac{\text{str}(u) \text{str}(v)}{2M} \right] \delta(c_u, c_v) \right).$$

Again, the weighted modularity of the network, Q_w , is the greatest of modularities obtained in all possible partitions of the network.