SPRAWOZDANIE
Analiza języka naturalnego, Ćwiczenie 2
Marcel Cielinski, Kornel Romański

1. **Miary bazujące na wektorach semantyki dystrybucyjnej**
   a. **Wybrane miary**

Miara kosinusowa - najpopularniejsza miara podobieństwa dla reprezentacji wektorowej tekstów, prezentująca wartość kosinusa kąta między dwoma wektorami. Wyższa wartość miary oznacza wyższe podobieństwo tekstów. Zgodnie z tą miarą podobieństwo jest definiowane jako:

$$sim(v_1, v_2) = \frac{v_1 * v_2}{|v_1||v_2|}$$

Odległość euklidesowa - klasyczna miara odległości pomiędzy wektorami w przestrzenii euklidesowej. Ponieważ wyższe podobieństwo oznacza niższą odległość między wektorami, to w tym przypadku teksty są bardziej podobne kiedy wartość miary jest niższa. Odległość euklidesowa definiowana jest jako:

$$dist(v_1,\ v_2) = \sqrt{\sum_i (v_{1_i} - v_{2_i})^2}$$

   b. **Przygotowanie wektorów osadzeń**

Do obliczania miar wykorzystane zostały gotowe wektory osadzeń otrzymane za pomocą modelu *FastText* wytrenowanego na korpusie *KGR10* metodami *skipgram* i *cbow* [1]. Ponieważ metody te bazują na obliczaniu reprezentacji ukrytych dla danego słowa na podstawie kontekstu, w którym ono występuje (sąsiednie słowa w tekście), a dwa słowa występujące w tym samym kontekście są powiązane, ale nie muszą być podobne, to powyższe miary między wektorami będą służyły do określania powiązania słów, a nie ich podobieństwa.

2. **Miary bazujące na sieci WordNet**
   a. **Wybrane miary** [2]

Miara WuPalmer - oblicza ona pokrewieństwo, biorąc pod uwagę głębokość obu synsetów w taksonomiach *WordNet*, wraz z głębokością *LCS (Least Common Subsumer)*. Otrzymane w ten sposób podobieństwa oparte są o miejsce występowania względem siebie

synsetów w drzewie zbudowanym na podstawie hiperonimii. Wartości z przedziału (0,1].

$$wupalmer(n_1, n_2) = \frac{2 \cdot depth(lcs(n_1,n_2))}{depth(n_1) + depth(n_2) + 2 \cdot depth(lcs(n_1,n_2))}$$

, gdzie:

- $depth$ - głębokość w hierarchii grafu
- lcs - Least Common Subsumer - najniższy wspólny przodek

Miara LeacockChodorow - miara ta zlicza liczbę krawędzi pomiędzy synsetami w hierarchii *WordNet*. Wartość ta jest następnie skalowana przez maksymalną głębokość hierarchii. Wartość korelacji uzyskuje się przez przyjęcie ujemnego logarytmu naturalnego tej skalowanej wartości.

$$leacockchodowow(n_1, n_2) = -\log \frac{dist(n_1,n_2)}{2 \times max_{c \in wordnet}\, depth(c)}$$

, gdzie:

- $dist$ - odległość między węzłami w grafie
- $depth$ - głębokość w hierarchii grafu

b. Przygotowanie sieci WordNet

Wykorzystana została polska słowosieć [3], na podstawie której zbudowano graf zgodnie z poniższymi krokami:

- jako wierzchołki użyto synsetów,
- krawędzie ograniczono do skierowanych relacji hiperonimii między synsetami,
- usunięto dwa występujące cykle (wierzchołki w cyklu połączono w jeden wierzchołek),
- aby otrzymać graf o charaktere drzewa dodano również wierzchołek spełniający rolę korzenia, który połączono do wszystkich wierzchołków w grafie nie mających dotychczas żadnych krawędzi wchodzących.

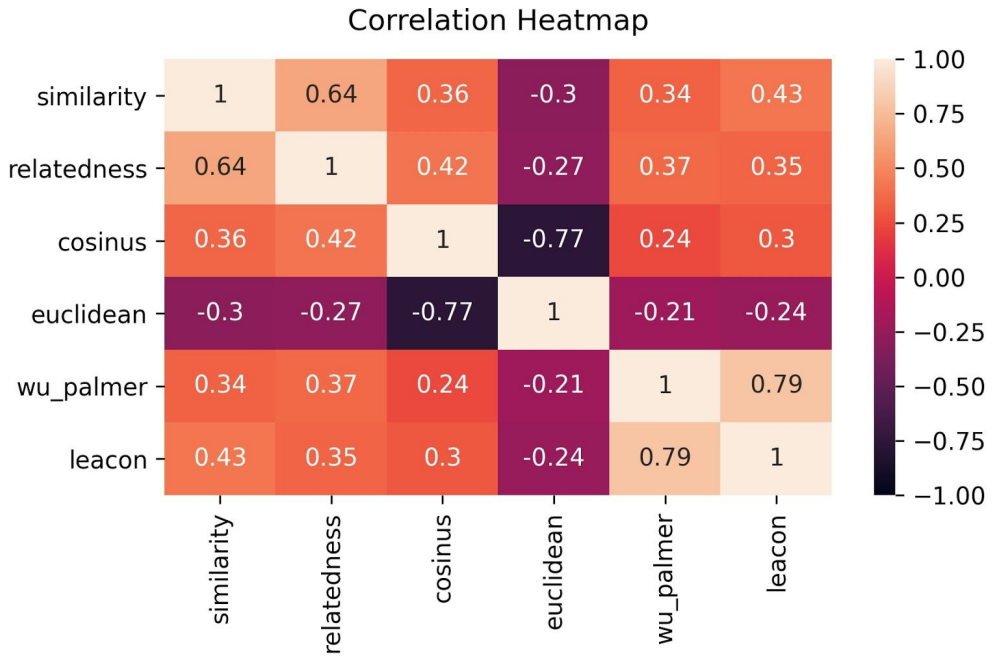
3. Wyniki

a. Wyniki dla zbioru SimLex [4]

Do porównania otrzymanych wyników z danymi zbioru *SimLex* wykorzystano korelację pearsona.

Poniżej przedstawiono macierz wszystkich korelacji pomiędzy kolumnami *similarity* i *relatedness* w zbiorze *SimLex* oraz wynikami metryk dla par słów w tym zbiorze. Należy zwrócić uwagę na występowanie ujemnych korelacji, które wynikają z tego, że odległość

euklidesowa oznacza większe podobieństwo tekstów kiedy jest mniejsza. Stąd w trakcie porównywania metryk rozpatrywane są wartości bezwzględne korelacji.



W przypadku miar opartych na wektorach semantyki dystrybucyjnej wyniki porównujemy z kolumną *relatedness*, dlatego interesujące nas korelacje to:

- **miara kosinusowa:** **0.42**
- **odl. euklidesowa:** **0.27**

Miary bazujące na słowosieci porównujemy z kolumną *similarity*:

- **miara WuPalmer:** **0.34**
- **miara LeacockChodorow:** **0.43**

b. Najbardziej podobne słowa w zadanym słowniku

Implementacja umożliwia wygenerowania listy *k* najbardziej podobnych słów według wybranej miary dla słów z ustalonego słownika.

Przeprowadzono eksperyment na wybranym słowniku, składającym się z następujących 10 słów:

- {'sufit', 'pochłonąć', 'okoliczność', 'rubież', 'upraszać', 'rytm', 'nowoczesny', 'pojemnik', 'gwałtowny', 'pudełko'}

W poniższej tabeli przedstawiono po 3 zestawy najbardziej podobnych słów zadanego słownika, z podziałem na stosowaną metrykę oraz

wartość *k*. W ostatniej kolumnie umieszczono wskazania miar dla każdego z zestawów.

METRYKA	*k*	SŁOWA	WARTOŚĆ
Cosinus	2	1. 'pojemnik', 'pudełko' 2. 'pochłonąć', 'gwałtowny' 3. 'okoliczność', 'gwałtowny'	0.6545 0.6520 0.5974
Cosinus	3	1. 'pochłonąć', 'upraszać', 'gwałtowny' 2. 'pochłonąć', 'nowoczesny', 'gwałtowny' 3. 'pochłonąć', 'okoliczność', 'gwałtowny'	0.5812 0.5589 0.5588
Euclidean	2	1. 'pochłonąć', 'gwałtowny' 2. 'pojemnik', 'pudełko' 3. 'pochłonąć', 'upraszać'	21.4629 21.7708 23.3221
Euclidean	3	1. 'pochłonąć', 'upraszać', 'gwałtowny' 2. 'pochłonąć', 'rubież', 'gwałtowny' 3. 'pochłonąć', 'okoliczność', 'gwałtowny'	23.5361 24.1262 24.1590
WuPalmer	2	1. 'sufit', 'rubież' 2. 'rytm', 'pojemnik' 3. 'okoliczność', 'nowoczesny'	0.4615 0.2857 0.2000
WuPalmer	3	1. 'sufit', 'okoliczność', 'rubież' 2. 'sufit', 'rubież', 'nowoczesny' 3. 'sufit', 'rubież', 'pojemnik'	0.2751 0.2751 0.2751
LeaChodorow	2	1. 'sufit', 'rubież' 2. 'sufit', 'okoliczność' 3. 'sufit', 'pudełko'	2.4159 2.2336 2.2336
LeaChodorow	3	1. 'sufit', 'rubież', 'pudełko' 2. 'sufit', 'okoliczność', 'rubież' 3. 'sufit', 'okoliczność', 'pudełko'	2.2944 2.2430 2.1822

c. Wnioski

- Miary LeacockChodorow i kosinusowa dały wyniki najbardziej zbliżone do anotacji w zbiorze *SimLex*.
- Najgorzej sprawdziła się odległość euklidesowa, która znacznie odstaje od konkurencyjnej miary kosinusowej.
- Miary oparte na wektorach osadzeń zwróciły lepiej interpretowalne wyniki w punkcie b.
- Metryki oparte na słowosieci wymagają znacznie więcej czasu do ich obliczenia, w szczególności metryka WuPalmer.

4. Źródła

1. https://clarin-pl.eu/dspace/handle/11321/606
2. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*, A. Budanitsky, G. Hirst
3. http://plwordnet.pwr.wroc.pl/wordnet/
4. *SimLex-999 for Polish*, A. Mykowiecka, M. Marciniak, P. Rychlik