Analisis Hasil Penerapan Metode *Distributional Semantic* untuk Kesamaan Semantik pada Bahasa Indonesia

Muhammad Taufik Wahdiat[1], Ade Romadhony, S.T.,M.T.[2], Said Al Faraby, S.T.,M.Sc.[3]

[1,2,3,]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung
[1]wahdiat@gmail.com, [2]ade.romadhony@telkomuniversity.ac.id, [3]saidalfaraby@telkomuniversity.ac.id

Abstrak

Kesamaan semantik adalah metrik kesamaan antar kata, kalimat atau dokumen yang berbagi dalam elemen makna. Perhitungan terkaitan semantik memiliki peranan penting dalam *data mining,* pengambilan informasi, dan bahkan *natural language processing.* Pada bahasa Indonesia, perhitungan kesamaan semantik mendapat peran penting karena banyak dimanfaatkan untuk aplikasi lain, seperti klasifikasi teks. Pengukuran kesamaan semantik dapat dilakukan dengan pendekatan berbasis korpus dan pendekatan berbasis kamus. Pada Tugas Akhir ini dilakukan pembangunan model kesamaan semantik berbasis korpus yang direpresentasikan dengan distributional *semantic vector*. Model kemudian diujikan pada beberapa pasang kata dengan derajat kesamaan semantik bervariasi. Model kesamaan semantik dibangun berdasar korpus Wikipedia Bahasa Indonesia, dengan metode *word2vec*. Hasil pengujian pada dataset uji yang juga digunakan pada penelitian sebelumnya berdasar pada referensi *SimLex999* dan *Rubenstein-goodenough* menunjukkan nilai korelasi yang diperoleh 0.2753. Walaupun nilai korelasi tersebut lebih kecil dibanding nilai pada penelitian sebelumnya dengan pendekatan korpus, terdapat beberapa kasus di mana model semantik berbasis korpus mampu menangkap korelasi semantik lebih baik.

Kata kunci : kesamaan semantik, bahasa Indonesia, persamaan kosinus.

Abstract

Semantic similarity is similarity metric between words, sentences or documents that shares element of meaning. Semantic similarity measurement has important role in data mining, information retrieval and even natural language processing. In Indonesian language, semantic similarity measurement has important role because it is widely used for other application, such as text classification. Semantic similarity can be done by corpus based approach and dictionary based approach. In this thesis, the development of corpus based semantic similarity model is represented by distributional semantic vector. The model is then tested on several pairs of words with varying degrees of semantic similarity. The semantic similarity model was build based on Indonesian Wikipedia corpus, with word2vec method. The test result on test dataset which used in previous studies based on SimLex999 dan Rubenstein-goodenough references show the correlation value obtained is 0.2753. Although the correlation value is smaller than value in previous study with the corpus approach, there are numbers of cases where the corpus based semantic model is able to capture the semantic correlation better.

Keywords: semantic similarity, Indonesian language, cosinus similarity.

1. Pendahuluan

Bahasa digunakan sebagai sarana komunikasi antar sesama manusia. Sebagai bangsa Indonesia, Bahasa Indonesia digunakan sebagai bahasa nasional untuk mempersatukan berbaagai bahasa daerah yang ada di Indonesia. Perkembangan yang semakin melesat menyebabkan ada beberapa kata bahasa Indonesia yang mengalami baik perubahan maupun penambahan kata. Beberapa kata yang berbeda dapat memiliki satu arti yang sama walaupun memiliki perbedaan dalam tulisan maupun pengucapan. Untuk mengukur persamaan antar kata ini dapat digunakan pengukuran kesamaan semantik teks.

Pengukuran kesamaan semantik teks terdiri dari perhitungan kesamaan antar istilah, pernyataan maupun teks yang memiliki arti mirip namun tidak memiliki kesamaan dalam pengurutan kata atau huruf. Hal ini merupakan permasalahan penting dalam perbidangan yang terkait dengan komputer, misalnya dalam pengambilan informasi [1]. Metode berbasis vektor merupakan salah satu metode yang dapat digunakan dalam pengukuran kesamaan semantik antar kata. Metode berbasis vektor termasuk ke dalam *distributional semantic*. *Distributional semantic* menghitung kesamaan semantik teks dengan cara merepresentasikan jumlah kemunculan kata dalam dokumen sebagai vektor, yang kemudian dihitung jarak antar vektor masing-masing menggunakan nilai kosinus.

Pada tugas akhir ini akan diimplementasikan perhitungan kesamaan semantik bahasa Indonesia berbasis korpus. Menggunakan Gensim yang merupakan *library* pada bahasa pemrograman Pyhton untuk mengimplementasikan algoritma *word2vec*, secara otomatis menemukan struktur semantik dengan memeriksa pola kemunculan statistik dalam dokumen yang dilatih pada korpus. Korpus yang dipakai pada gensim diambil

dari *text dump* Wikipedia bahasa Indonesia. Untuk mengevaluasi sistem yang dibuat, maka nilai hasil kesamaan semantik yang didapat selanjutnya dibandingkan dengan penelitian serupa yang menggunakan basis pengetahuan KBBI. Dataset yang digunakan berdasarkan referensi dari *Simlex999* dan *Rubenstein-goodenough* dan *gold standard* yang didapat berdasarkan hasil kuisioner terhadap 31 orang.

2. Studi Terkait

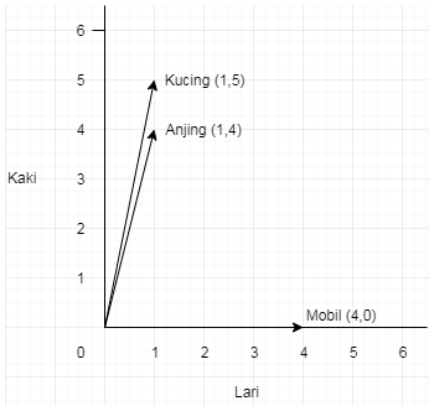
2.1. Distributional Semantic

Distributional hypothesis menyatakan "kemunculan kata pada konteks yang mirip cenderung memiliki arti yang mirip" [2]. *Distributional Semantic* terealisasikan dari *distributional hypothesis* tersebut. Dalam *Distributional Semantic Model*, setiap kata terepresentasikan oleh vektor matematis yang berupa urutan angka menggunakan statistik berbasis korpus. Nilai dalam komponen vektor yang dihasilkan sebagian besar berupa jumlah kemunculan suatu kata dan kata lain didekatnya dalam sebuah konteks. Alur pembangunan model dari *distributional semantic* dimulai dengan *preprocess* korpus. Lalu pengumpulan metriks berdasarkan kemunculan kata yang hasilnya digunakan untuk menghitung kesamaan antar kata. Sebagai contoh, tabel di bawah adalah konteks kata yang dijadikan vektor.

	Lari	Kaki
Anjing	1	4
Kucing	1	5
Mobil	4	0

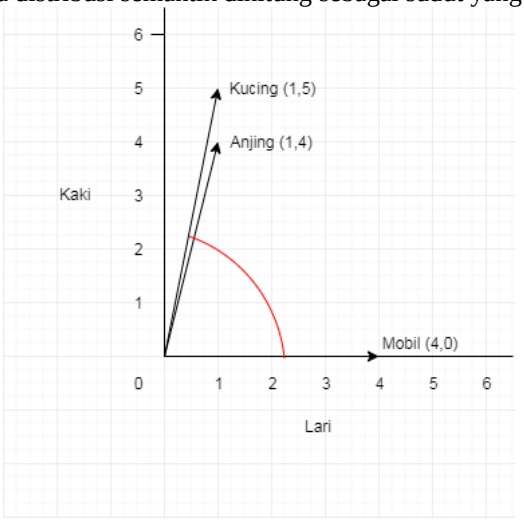
Tabel 1 Contoh Konteks Kata

Dari konteks kata tersebut, kemudian dapat dibuat vektor.



Gambar 1 Vektor Contoh Distribusi Semantik

Nilai kesamaan semantik pada distribusi semantik dihitung sebagai sudut yang terbentuk di antara vektor.



Gambar 2 Sudut dalam Vektor Distribusi Semantik

2.2 Korelasi Pearson

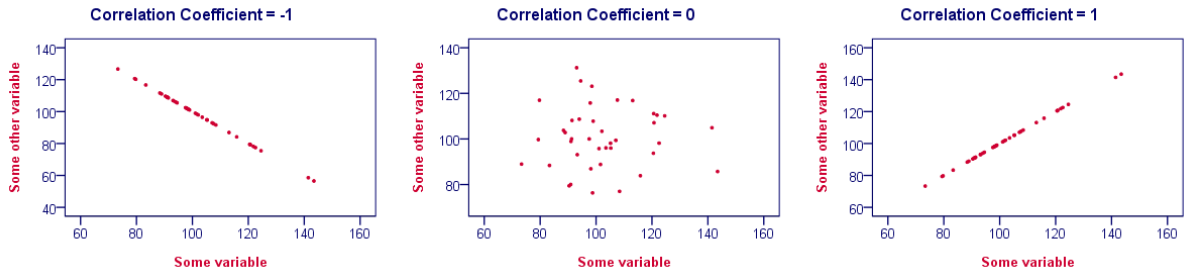
Korelasi pearson merupakan nilai di antara -1 dan 1 yang menunjukkan sejauh mana dua variabel terkait secara linear. Korelasi pearson hanya cocok untuk variabel metriks. Berikut merupakan penjelasan dari nilai korelasi :

Korelasi yang bernilai -1 mengindikasikan kedua variabel berkorelasi secara linear negatif. Nilai korelasi tidak pernah lebih rendah dari -1

Korelasi yang bernilai 0 berarti kedua variabel tidak memiliki korelasi linear, namun berkemungkinan ada korelasi non-linear.

Korelasi yang bernilai 1 berarti kedua variabel berkorelasi secara linear positif. Nilai korelasi tidak pernah lebih dari 1

Gambar 1 di bawah merupakan korelasi pearson yang divisualisasikan sebagai scatterplot.



Gambar 3 Visualisasi Scatterplot Korelasi Pearson dengan Nilai -1, 0 dan 1

Formula dari korelasi Pearson antara variabel x dan y dihitung dengan rumus :

$$rXY = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

Dimana $\bar{X}$ dan $\bar{Y}$ merupakan nilai rata-rata dari array variabel x dan variabel y [6].

Berikut kriteria hubungan dari korelasi pada tabel 2 di bawah.

Nilai r	Kriteria Hubungan
0	Tidak ada korelasi
0 - 0,5	Korelasi lemah
0,5 – 0,8	Korelasi sedang
0,8 – 1	Korelasi kuat
1	Korelasi sempurna

Tabel 2 Kriteria Hubungan Korelasi

Sebagai contoh, untuk mengetahui hubungan nilai *gold standard* terhadap nilai kesamaan kata dan mengetahui keeratan hubungan antar kedua variabel, data sebanyak 10 buah sampel pasangan kata diambil. Nilai *gold standard* dan nilai kesamaan sampel dapat dilihat pada tabel 3 di bawah.

Sampel	Nilai *gold standard*	Nilai kesamaan
1	0.32	0.29
2	0.53	0.67
3	0.66	0.87
4	0.87	0.54
5	0.32	0.22
6	0.25	0.45
7	0.67	0.7
8	0.19	0.23
9	0.64	0.7
10	0.72	0.7

Tabel 3 Tabel Contoh Data Kesamaan

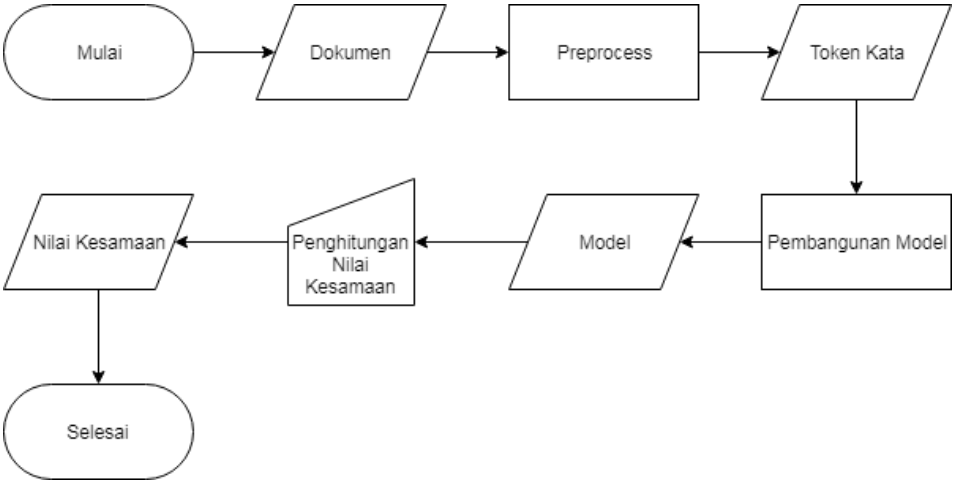
.

Dengan menggunakan rumus korelasi Pearson, nilai korelasi dari kesepuluh sampel di atas menghasilkan nilai r = 0.762745. Korelasi koefisien yang dihasilkan antara *gold standard* dengan nilai kesamaan kata memiliki nilai linear positif yang sedang.

3. Gambaran Umum Sistem

Untuk menghitung nilai kesamaan semantik digunakan Gensim sebagai sarana dalam bahasa pemrograman Python. Dimulai dari *input* dokumen untuk pembangunan model pada Gensim. Dokumen yang digunakan dalam perhitungan kesamaan semantik pada gensim adalah *text dump* Wikipedia yang memiliki ekstensi txt utf-8. *Input* dokumen yang ada kemudian dilakukan *preprocessing* untuk membuat model corpus yang berupa token kata. Token kata yang tercipta kemudian disimpan dalam model gensim. Langkah berikutnya yaitu *training* model untuk merepresentasikan kata sebagai vektor, yang digunakan untuk mengukur beban token kemunculan kata pada dokumen. Setelah dilakukan model *training*, selanjutnya dapat dihitung nilai dari kesamaan semantik kata. Hasil yang didapatkan dari perhitungan kesamaan semantik kemudian dipakai untuk menghitung korelasi pearson.

Alur dari keseluruhan sistem yang dibangun dapat dilihat pada gambar di bawah.



Gambar 4 Flowchart Penelitian yang Dilakukan

Dari flowchart pada gambar 4, gambaran sistem sebagai berikut :

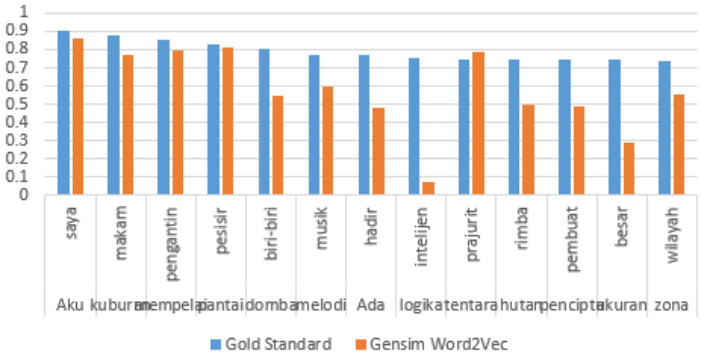
1. Dimulai dengan *input* dokumen, di mana pada penelitian ini berupa Wikipedia *text dump* bertipe txt file utf-8.
2. Langkah berikutnya yaitu melakukan *preprocessing*, di mana pada langkah ini dilakukan penghilangan kata *stopword*, pengubahan menjadi huruf kecil dan pemisahan masing-masing kata yang kemudian didapatkan token masing-masing kata.
3. Selanjutnya dilakukan pembangunan model menggunakan *word2vec* yang telah disediakan dari gensim. Parameter yang dipakai untuk membangun model yaitu kemunculan kata minimal 2, mengabaikan semua kemunculan kata yang berjumlah 1.
4. Setelah model dibangun, selanjutnya dapat dihitung nilai dari kesamaan kata. Pada penelitian ini penghitungan nilai kesamaan dilakukan secara manual terhadap setiap pasangan kata.

Nilai kesamaan kata yang didapatkan berjumlah 171 kata dari total 180 kata pada penelitian perbandingan, penghilangan 9 kata dilakukan karena kata tersebut tidak terdapat pada model yang dibuat dari dokumen Wikipedia *text dump* terpilih. Hasil dari nilai kesamaan kata selanjutnya dihitung nilai selisihnya dengan nilai gold standard untuk membandingkan perhitungan kesamaan kata dengan penelitian pembanding. Nilai kesamaan kata lalu dapet digunakan bersama dengan nilai gold standard untuk menghitung nilai korelasi Pearson, mencari hubungan antar variabel dan dibandingkan agar mengetahui nilai korelasi yang terbaik.

4. Evaluasi

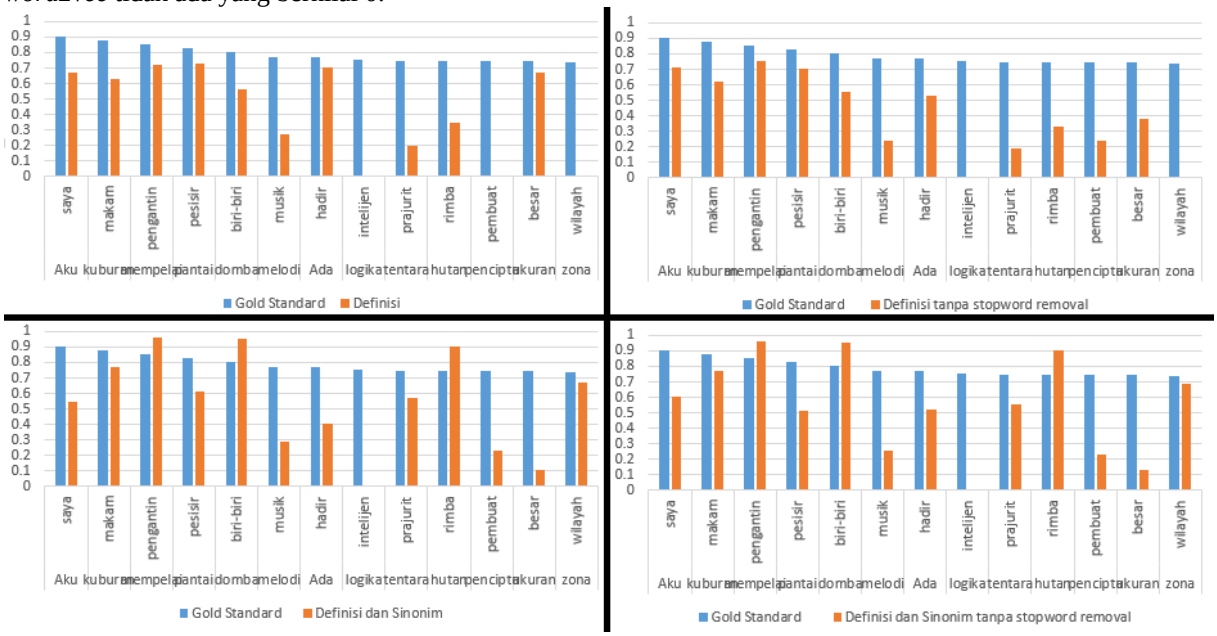
4.1 Hasil Pengujian

Gambar di bawah merupakan cuplikan hasil perhitungan nilai kesamaan semantik yang dihasilkan dari gensim, gold standard, dan empat hasil perhitungan kesamaan semantik menggunakan basis pengetahuan KBBI dan API Kateglo yang dijadikan perbandingan. Cuplikan diambil dengan 13 nilai kesamaan tertinggi pada gold standard.



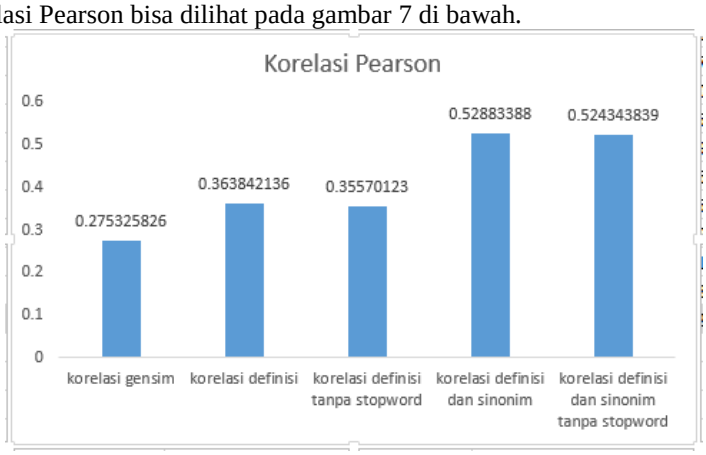
Gambar 5 Cuplikan Perbandingan Nilai Kesamaan Semantik antara Gold Standard dan Gensim Word2Vec

Gambar 5 di atas merupakan cuplikan dari nilai kesamaan antara gold standard dan nilai kesamaaan yang dihasilkan pada penelitian ini. Dapat dilihat nilai kesamaan kata yang dihasilkan rata-rata berbeda tidak jauh dengan nilai kesamaan gold standard yang ada. Dari keseluruhan nilai yang ada baik pada gold standard dan hasil dari *word2vec* tidak ada yang bernilai 0.



Gambar 6 Cuplikan Perbandingan Nilai Kesamaan Semantik antara Gold Standard dan Penelitian Pembanding

Pada gambar 6 merupakan cuplikan nilai gold standard dengan penelitian pembanding yang memiliki empat percobaan. Dari gambar 6 dapat dilihat bahwa pada setiap penelitian pembanding terdapat nilai kesamaan yang bernilai 0. Nilai dari korelasi Pearson bisa dilihat pada gambar 7 di bawah.



Gambar 7 hasil nilai korelasi pearson dari tiap perhitungan

	Nilai Rata-Rata Selisih Kesamaan Semantik Gold Standard dengan Nilai Lainnya
Gensim Word2Vec	0.1942
Definisi	0.3296
Definisi tanpa Stopword Removal	0.3259
Definisi dan Sinonim	0.2727
Definisi dan Sinonim tanpa Stopword Removal	0.2709

Tabel 4 Hasil Perhitungan Selisih Nilai Rata-Rata Gold Standard dengan Nilai Sistem

4.2 Analisis Hasil Pengujian

Bisa dilihat dari gambar 5 dan gambar 6 di atas, hasil perhitungan kesamaan semantik pada gensim word2vec rata-rata memiliki nilai kesamaan yang lebih dekat dengan *gold standard*, dibandingan dengan melalui pembobotan tf-idf. Namun hasil pada perhitungan korelasi pearson pada gambar 7 menunjukkan nilai pada gensim *word2vec* paling kecil dibandingkan dengan nilai lainnya. Dengan nilai 0.2753, dapat diambil kesimpulan bahwa nilai korelasi dari gensim *word2vec* dan *gold standard* memiliki kesamaan linear yang positif, namun mendekati angka 0 daripada 1 yang berarti kesamaan linear positif yang lemah. Pada tabel 4 dapat dilihat hasil nilai selisih dari *gold standard* dengan nilai dari sistem. Nilai kesamaan semantik yang terbaik adalah nilai dari gensim *word2vec*, di mana nilai selisih yang didapatkan selisih paling kecil dari semua percobaan, yaitu bernilai 0.1942.

5. Kesimpulan

Berdasarkan hasil pengujian yang telah dilakukan, dapat diambil kesimpulan bahwa nilai korelasi terbaik yang didapatkan adalah 0.5288, nilai korelasi antara *gold standard* dan pengujian melalui pembobotan tf-idf definisi dan sinonim. Nilai korelasi pada tugas akhir ini bernilai 0.2753 antara *gold standard* dan gensim *word2vec*, nilai korelasi terkecil dari semua perbandingan. Parameter terbaik yang mempengaruhi nilai kesamaan semantik yaitu 0.1942, nilai selisih kesamaan semantik antara gensim *word2vec* dengan *gold standard*.

Saran yang dapat dijadikan bahan penelitian untuk dijadikan pengembangan tugas akhir ini adalah :

- Menggunakan ukuran korpus yang lebih besar dari 245 MB, ukuran korpus yang digunakan pada tugas akhir ini, sehingga kata yang dapat dihitung nilai kesamaan semantik lebih lengkap dan beragam.
- Melakukan *preproccess* yang lebih lengkap dalam memproses korpus karena hanya dilakukan *preproccess* sederhana pada tugas akhir ini.
- Mencari parameter terbaik dalam pembangunan model *word2vec* agar didapatkan nilai kesamaan semantik terbaik.

Daftar Pustaka

[1] Jorge Martinez-Gil. An overview of textual semantic similarity measures based on web intelligence. Artificial Intelligence Review, Springer Verlag, 2012, 42 (4), pp.935-943. 10.1007/s10462-012-9349-8 . hal-01630890

[2] Kiela, Douwe, and Clark, Stephen. *A Systematic Study of Semantic Vector Space Model Parameters. 2014.*

[3] Harispe, Sebastien, et al. "Semantic Similarity from Natural Language and Ontology Analysis." *Systhesis Lectures on Human Language Technologies,* vol. 8, no. 1, 2015.

[4] "Polyglot." *Rami Al-Rfou.* (https://sites.google.com/site/rmyeid/projects/polyglot).

[5] "Gensim: Topic Modelling for Humans." *Radim*

[6]SPSS Tutorials. (https://www.spss-tutorials.com/pearson-correlation-coefficient/).

[7] Jurafsky, Dan, and Martin, James. *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* third ed., Dorling Kindersley Pvt, Ltd., 2014.

[8] Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2013. Frege in Space: A program for compositional distributional Semantics. *Linguistic Issues in Language Technologies (LiLT).*