Analisis dan Implementasi Kesamaan Semantik Antar Kata Berbahasa Inggris Menggunakan *Pointwise Mutual Information Max* dengan Wikipedia Sebagai *Corpus*

Shervano Naodias Siagian[1], **Moch. Arif Bijaksana**[2]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]S1 Teknik Informatika
[1]Shervano.Naodias@gmail.com, [2]arifbijaksana@telkomuniversity.ac.id

Abstrak

Sejumlah besar data informasi dapat disimpan dalam basis data di internet. Salah satu jenis data informasi yang sering digunakan oleh manusia adalah data dalam bentuk teks. Selama menggunakan data teks untuk mencari sesuatu di internet biasanya memanfaatkan kata kunci, padahal satu kata bisa saja memiliki makna yang berbeda. Misalnya mencari kata *"bat"* di internet, bisa saja *search engine* akan menampilkan artikel tentang kelelawar atau alat pemukul *baseball*. Seiring adanya kasus tersebut memicu penelitian yang terkait data teks meningkat, penelitian ini sering disebut dengan *text mining*. Salah satu implementasi dari penelitian data teks adalah *semantic similarity* yaitu melihat kemiripan makna pasangan kata dengan memberikan nilai *similarity*. Untuk menganalisis kemiripan makna pada pasangan kata diperlukan adanya suatu sistem yang dapat menghitung nilai kemiripan antara sepasang kata dengan menggunakan metode PMI*max* dan menggunakan *gold standard* untuk mendapatkan nilai korelasi sistem sebagai evaluasi. Dari hasil penelitian menggunakan korelasi *pearson*, didapat nilai korelasi terbesar yaitu 0.71 (Miller-Charles) dan nilai korelasi terkecil adalah -0.03 (SimLex-999). Hal tersebut disebabkan karena banyak pasangan kata yang ada didalam *gold standard (Miller-Charles)* ada didalam korpus wikipedia sehingga memiliki nilai *similarity*. Sedangkan untuk *gold standard (SimLex-999)* banyak pasangan kata yang tidak ada didalam korpus wikipedia, sehingga nilai *similarity* menghasilkan nilai 0.

Kata kunci : ***Semantic Similarity, Pointwise Mutual Information Max, Wikipedia, Gold Standard, Pearson Correlation***

Abstract

Majority of data information stored in the internet nowadays are based on texts, this system ease the internet users to search with keywords they attemp for search engine, e.g. word "bat" would probably show up as animal or in the otherhand would probably show up as with the article about baseball. Over the existence of such cases trigger has also increased the text data related research, which is often referred to as text mining. One of the implementations on the data text research is the semantic similarity, that by giving couples the value of the similarity to measure the resemblance of the meaning of the word-pairs. To analyze the similarity of meaning to the word-pairs, a system was built that can calculate the value of the similarity between a pair of words semantically using *Pointwise Mutual Information Max* (PMI*max*) and use the *gold standard* to evaluate and to get the value of the correlation system. The result of PMI*max* method, obtained the largest correlation value i.e. 0.71 with *gold standard* dataset *Miller-Charles* use *Pearson Correlation*. While the smallest correlation value is -0.03 with the *gold standard* dataset of (*SimLex-999*). This is because many of the *gold standard* dataset *Miller-Charles* word-pairs were also be found in the wikipedia corpus. consequently, these word-pairs will obtain their similarity value. Consequently, for those word-pairs that were of *gold standard* dataset *SimLex-999* but they were not be in the wikipedia corpus, they will produces a similarity value of 0.

Keywords: ***Semantic Similarity, Pointwise Mutual Information Max, Wikipedia, Gold Standard, Pearson Correlation***

1. Pendahuluan

Latar Belakang

Sejumlah besar data informasi dapat disimpan dalam basis data di internet [1]. Salah satu jenis data informasi yang sering digunakan oleh manusia adalah data dalam bentuk teks. Selama menggunakan data teks untuk mencari sesuatu di internet biasanya memanfaatkan kata kunci, padahal satu kata bisa saja memiliki kesamaan atau makna

yang berbeda. Misalnya kita mencari teks di internet dengan menggunakan kata kunci *"bat"*, bisa saja *search engine* menampilkan artikel yang membahas kelelawar dan bisa saja membahas tentang alat pemukul *baseball*. Seiring adanya kasus tersebut memicu penelitian yang terkait data teks juga meningkat, penelitian ini sering disebut dengan *text mining*. Banyak implementasi dari penelitian data teks tersebut yang digunakan, salah satunya adalah *semantic similarity*.

Semantic Similarity, merupakan salah satu *task* pada *text mining* dalam ranah *Natural Language Processing* untuk melihat kemiripan makna pasangan teks dengan memberikan nilai kemiripan antara sepasang kata secara semantik [2]. Banyak riset yang terkait yang telah dilakukan untuk mendapat nilai kemiripan antar teks. Pengukuran kemiripan kata yang telah banyak diketahui biasanya berdasarkan pada WordNet [3]. Banyak algoritma atau metode yang dapat diimplementasikan untuk pengukuran dari *Semantic Similarity* antara sepasang kata, misalnya *Salient Semantic Analysis* dan *Pointwise Mutual Information* (PMI). PMI telah banyak diketahui memiliki kecenderungan memberikan skor yang tinggi untuk sepasang kata [4]. Dalam tugas akhir ini, algoritma yang diimplementasikan adalah PMI*max*, yang merupakan modifikasi dari metode PMI [2]. Berbeda dengan PMI yang pengukurannya setiap kata memiliki hanya satu *sense*, sedangkan PMI*max* mengestimasi korelasi antara dua kata dan korelasi antara makna terdekat kedua kata tersebut karena setiap kata mungkin memiliki lebih dari satu makna.

Pada tugas akhir kali ini akan dilakukan pencarian nilai *semantic similarity* untuk kata Bahasa Inggris menggunakan metode PMI*max*. Dalam pencarian nilai tersebut, *input*an yang digunakan adalah kumpulan artikel yang sudah digabungkan menjadi satu buah korpus Bahasa Inggris. Proses pengumpulan artikel akan memanfaatkan wikipedia, yang merupakan salah satu sumber korpus multibahasa yang besar dan dapat diakses secara bebas [5]. Untuk melihat tingkat akurasi dari metode yang digunakan, akan dilakukan pencarian nilai korelasi antara nilai *semantic similarity* yang dihasilkan dengan menggunakan metode PMI*max* dengan nilai yang ada pada *gold standard*. Ada beberapa jenis korelasi yang dapat digunakan yaitu korelasi *pearson* dan korelasi *spearman*. Namun yang digunakan pada Tugas Akhir ini adalah korelasi *pearson* dan *gold standard* yang digunakan adalah *WordSim-353, SimLex-999,* dan *Miller-Charles*. Diharapkan dengan dilakukannya tugas akhir ini maka dapat diketahui hasil pengukuran *similarity* antara teks.

Topik dan Batasannya

Dari beberapa pemaparan masalah yang ada pada latar belakang, dapat diambil beberapa hal yang dapat dijadikan rumusan masalah yaitu sebagai berikut.

1. Bagaimana metode PMI*max* bekerja dalam melakukan perhitungan antara pasangan kata dalam Bahasa Inggris?
2. Bagaimana implementasi rancangan metode tersebut dilakukan?
3. Berapa nilai *similarity* yang dihasilkan dari penggunaan metode PMI*max* tersebut dan korelasinya berdasarkan dataset *gold standard* (*WordSim-353, SimLex-999,* dan *Miller-Charles*)?

Terdapat beberapa batasan masalah yang ada pada pengerjaan tugas akhir ini yaitu sebagai berikut.

1. Data yang digunakan adalah dataset dengan format *".txt"* yang berisi teks menggunakan Bahasa Inggris.
2. Korpus yang digunakan mengambil *knowledge based* yang bersumber dari Wikipedia Bahasa Inggris dan ukuran korpus yang digunakan maksimal 1MB.
3. Dataset *gold standard* yang digunakan pada tugas akhir ini adalah *WordSim-353, SimLex-999,* dan *Miller-Charles* sebagai pembanding nilai *Semantic Similarity*.
4. Menggunakan korelasi *pearson* untuk menghitung akurasi sistem.
5. Kata yang dapat dimasukan ke sistem hanya pasangan kata yang terdapat pada dataset *gold standard* yaitu *WordSim-353, SimLex-999,* dan *Miller-Charles*.

Tujuan

Tujuan dari tugas akhir ini diantaranya :

1. Mengimplementasikan perhitungan nilai *semantic similarity* dengan menggunakan metode PMI*max*
2. Menganalisis hasil implementasi perhitungan nilai *semantic similarity* dengan menggunakan metode PMI*max*
3. Mengetahui dan mencari nilai korelasi dari hasil implementasi perhitungan nilai *semantic similarity* dengan menggunakan korelasi *pearson* dan metode PMI*max* yang dibandingkan dengan *gold standard* yang sudah ditentukan yaitu *WordSim-353, SimLex-999,* dan *Miller-Charles*.

Organisasi Tulisan

Laporan penulisan tugas akhir ini disusun dengan sistematika sebagai berikut. Pertama Studi Terkait, pada bagian ini berisi teori-teori dan literatur terkait untuk mendukung pengerjaan tugas akhir. Kedua Sistem yang Dibangun, pada bagian ini berisi tentang proses rancangan dari sistem atau produk yang dihasilkan dari tugas akhir. Ketiga Evaluasi, pada bagian ini berisi tentang hasil pengujian dan analisis hasil pengujian tugas akhir. Keempat Kesimpulan, pada bagian ini berisikan kesimpulan dan saran dari hasil tugas akhir.

2. Studi Terkait

2.1 *Semantic Similarity*

Semantic Similarity atau kesamaan semantik adalah suatu metode untuk menghitung suatu kemiripan dari makna, atau menghitung jarak suatu makna antara dua konsep berdasarkan ontologi yang telah diberikan [6]. Terdapat beberapa kategori perhitungan dalam *Semantic Similarity*, salah satunya adalah *Information Content* yang merupakan perhitungan dengan mencari banyaknya kemunculan *terms* dari suatu dokumen [6]. Pada tugas akhir ini, akan menggunakan *Information Content* karena akan dihitung banyaknya kemunculan dua *term* pada suatu dokumen.

2.2 *Pointwise Mutual Information Max (PMImax)*

Pointwise Mutual Information Max (PMI*max*) merupakan modifikasi dari PMI [2]. PMI*max* mengestimasi nilai maksimum korelasi antara dua kata, yaitu korelasi antara kedua makna terdekat [2]. Untuk melihat korelasi antara kedua makna terdekat pada metode ini, maka akan dicari nilai dari makna pada setiap kata.

$$PMI_{max}(w_1,w_2) = \log\left(\frac{\left(fd(w_1,w_2) - \frac{e^k}{N}\cdot\left(fw_1\cdot fw_2 - \frac{fw_1}{yw_1}\cdot\frac{fw_2}{yw_2}\right)\right)\cdot N}{\frac{fw_1}{yw_1}\cdot\frac{fw_2}{yw_2}}\right) \tag{1}$$

Persamaan (1), merupakan rumus dari metode PMI*max*. Dimana $fd(w_1,w_2)$ merupakan nilai kemunculan kata 1 (w_1) dan kata 2 (w_2) secara bersamaan dalam sebuah korpus [7], e^k merupakan konstansa dengan nilai tetapan 10, N merupakan total jumlah kata pada korpus, fw_1 dan fw_2 merupakan kemunculan kata 1 dan kata 2 secara individu, sedangkan yw_1 dan yw_2 merupakan nilai makna (*sense*) dari kata 1 (w_1) dan kata 2 (w_2). Pencarian nilai *sense*, dapat dilihat pada persamaan (2).

$$yw = \frac{(\log(fw)+q)^p}{(\log(700)+q)^p} \tag{2}$$

Dimana nilai q dan nilai p merupakan sebuah variabel. q adalah variabel yang memiliki *range* [-6 - 10] berkelipatan 1. Sedangkan variabel p memiliki *range* [0 - 10] berkelipatan 0,5.

2.3 *Matriks Evaluasi*

Matriks Evaluasi bertujuan untuk mengevaluasi kinerja dari suatu sistem yang telah dibangun. Salah satu cara untuk mengevaluasi sistem adalah dengan menghitung korelasi. Korelasi merupakan salah satu istilah statistik berupa nilai yang menyatakan hubungan antara dua variabel [8]. Nilai korelasi memiliki rentang $-1 \leq 0 \leq 1$. Sehingga nilai korelasi dapat dibagi kedalam tiga kelompok, yaitu korelasi positif (kedua variabel memiliki keterhubungan yang searah dan nilai korelasinya direntang (0,1]) [9], Tidak ada korelasi (kedua variabel memiliki keterhubungan yang tidak memiliki arah dan nilai korelasinya adalah 0) [9], dan Korelasi negatif (kedua variabel memiliki keterhubungan yang tidak searah dan nilai korelasinya berada direntang [-1,0)) [9]. Ada beberapa contoh perhitungan dalam menghitung nilai korelasi seperti korelasi *pearson* dan korelasi *spearman*, namun pada tugas akhir yang digunakan adalah korelasi *pearson*. Rumus korelasi *pearson* dapat dilihat pada persamaan (3).

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{3}$$

Dimana *n* merupakan jumlah pasangan kata, *x* merupakan nilai dari sistem, sedangkan *y* merupakan nilai dari *gold standard*.

2.4 *Wikipedia*

Wikipedia merupakan sebuah situs aplikasi ensiklopedia online terbesar yang digunakan sebagai tempat pertukaran berbagai macam informasi yang disajikan dalam bentuk suatu artikel [10]. Wikipedia menggunakan *software open source* yang banyak digunakan oleh situs-situs berbasis ensiklopedia lainnya yang dapat di-*download* secara bebas di *https://dumps.wikimedia.org/enwiki/* [11]. Wikipedia menyediakan artikel yang mendukung pengetahuan dasar perhitungan keterikatan kata dengan lebih terstruktur dan cakupan katanya pun lebih kompleks [12].

Data *dump* wikipedia memiliki banyak sekali struktur teks yang tidak dibutuhkan dalam perhitungan sistem, oleh karena itu untuk menyederhanakannya, isi konten artikel wikipedia diambil secara online menggunakan *API (Application Programming Interface) Python* berdasarkan judul-judul yang sudah direkap dalam bentuk teks sebelumnya. Pada tugas akhir ini, artikel wikipedia digunakan sebagai korpus.

2.5 *Gold Standard*

Gold Standard merupakan suatu nilai yang digunakan untuk menganalisis korelasi hasil dari suatu sistem. Standar ini dibuat berdasarkan beberapa pihak manusia yang ahli dibidangnya untuk menilai *semantic similarity* sepasang konsep atau istilah dalam skala tertentu. Pada tugas akhir ini, dataset *WordSim-353, SimLex-999,* dan *Miller-Charles* dijadikan sebagai acuan *gold standard* yang berupa skor *semantic similarity* pasangan kata. Dataset *WordSim-353* terdiri dari dua dataset, yaitu untuk *relatedness* untuk acuan keterkaitan semantik dengan 252 pasangan kata, dan *similarity* untuk acuan kesamaan semantik dengan 203 pasangan kata [13]. Pada tugas akhir ini hanya menggunakan dataset *WordSim-353 (Similarity)* saja sebagai acuannya karena dataset tersebut sesuai dengan kasus pada tugas akhir ini. *SimLex-999* terdiri dari 999 pasangan kata didalamnya [14], dan *Miller-Charles* terdiri dari 30 pasangan kata [15]. Potongan isi dari dataset *WordSim-353 (Similarity), SimLex-999,* dan *Miller-Charles* dapat dilihat pada Tabel 1.

Tabel 1. **Potongan dataset *WordSim-353 (Similarity), SimLex-999,* dan *Miller-Charles***

WordSim-353 (Similarity)			*SimLex-999*			*Miller-Charles*		
Kata 1	**Kata 2**	**Skor**	**Kata 1**	**Kata 2**	**Skor**	**Kata 1**	**Kata 2**	**Skor**
Tiger	Cat	7.35	Old	New	1.58	Asylum	Madhouse	3.61
Tiger	Tiger	10	Smart	Intelligent	9.2	Bird	Cock	3.05
Plane	Car	5.77	Hard	Difficult	8.77	Bird	Crane	2.97

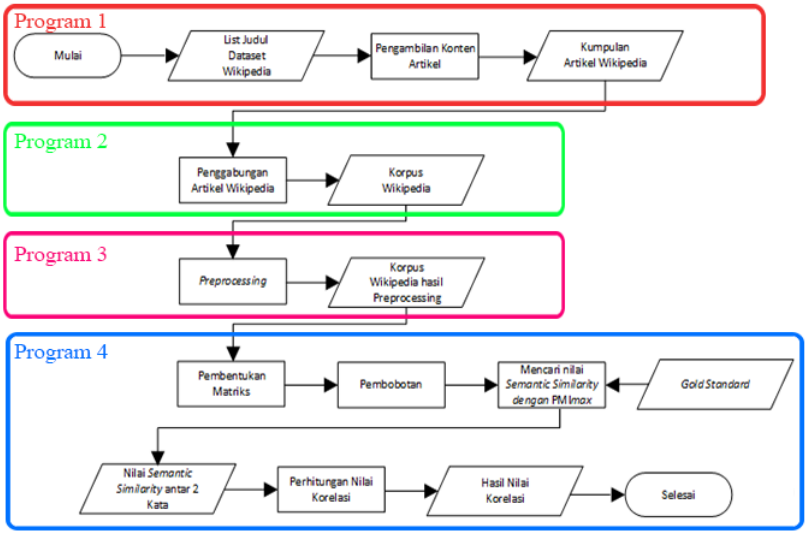
3. Sistem yang Dibangun

Sistem yang dibangun merupakan sistem berbasis desktop dengan menggunakan bahasa pemrograman *python* yang diprogram dalam aplikasi *anaconda*. Sistem terdiri dari empat program, yaitu program untuk mengambil artikel secara online, program untuk menggabungkan semua artikel menjadi satu buah korpus, program khusus untuk *preproscessing* korpus, dan program perhitungan PMI*max*. Data yang digunakan dalam perhitungan PMI*max* ada dua, yaitu data kumpulan artikel *dump* wikipedia sebagai *input*an dan data *gold standard* (*WordSim-353 (Similarity), SimLex-999,* dan *Miller-Charles*) sebagai pencarian nilai *semantic similarity*. Data-data tersebut dihitung secara semantik yang diolah menggunakan metode *Pointwise Mutual Information Max* (PMI*max*) untuk menentukan skor *similarity*. Kemudian *output* dari perhitungan tersebut berupa skor. Setelah skor *similarity* didapatkan, kemudian dihitung nilai korelasinya berdasarkan *gold standard* (*WordSim-353 (Similarity), SimLex-999,* dan *Miller-Charles*) dengan menggunakan persamaan *pearson*. Nilai korelasi menggambarkan seberapa efektif sistem dalam menjalankan fungsinya berdasarkan *gold standard* yang digunakan. Alur perancangan sistem dapat dilihat pada Gambar 1.

3.1 Pengambilan Konten Artikel

Tahap ini dilakukan khusus menggunakan program terpisah dengan program *preprocessing*, program penggabungan artikel dan program utama karena keterbatasan *resource* yang membutuhkan internet pada saat menjalankan program, sehingga dibutuhkan waktu lama dalam melakukan prosesnya. Proses ini membutuhkan data berupa file teks yang berisi daftar judul seluruh artikel wikipedia yang akan digunakan sebagai dataset acuannya. Daftar

judul tersebut direkap secara manual berdasarkan 30 judul atau 90 judul pertama data *dumb* wikipedia *enwiki-pages-meta-current1*. Setelah daftar judul di*input*kan, kemudian konten artikel diambil secara otomatis dengan menggunakan *API Python Wikipedia* secara berurutan. Program akan berhenti jika seluruh judul artikel sudah didapatkan isi kontennya yang berbentuk *xml*. Hasil data *xml* tersebut kemudian disimpan dengan merubah format datanya menjadi teks (*.txt*) dan otomatis direkap dalam satu folder sehingga menjadi dataset wikipedia mentah. Salah satu contoh isi konten *xml* wikipedia yang diambil secara online dapat dilihat pada Gambar 2.



Gambar 1. **Alur Perancangan Sistem**

b'<?xml version="1.0"?><api batchcomplete=""><warnings><extracts xml:space="preserve">"exlimit" was too large for a whole article extracts request, lowered to 1.</extracts></warnings><query><normalized><n from="automobile" to="Automobile" /></normalized><redirects><r from="Automobile" to="Car" /></redirects><pages><page _idx="13673345" pageid="13673345" ns="0" title="Car"><extract xml:space="preserve">A car (or automobile) is a wheeled motor vehicle used for transportation. Most definitions of car say they run primarily on roads, seat one to eight people, have four tires, and mainly transport

Gambar 2. **Potongan Data *XML* Wikipedia**

3.2 Penggabungan Artikel

Tahap ini menggabungkan 30 artikel atau 90 artikel wikipedia yang sudah didapat dari tahap pengambilan konten artikel menjadi satu buah korpus wikipedia yang nanti akan digunakan untuk tahap berikutnya. Contoh hasil dari penggabungan artikel wikipedia dapat dilihat pada Gambar 3.

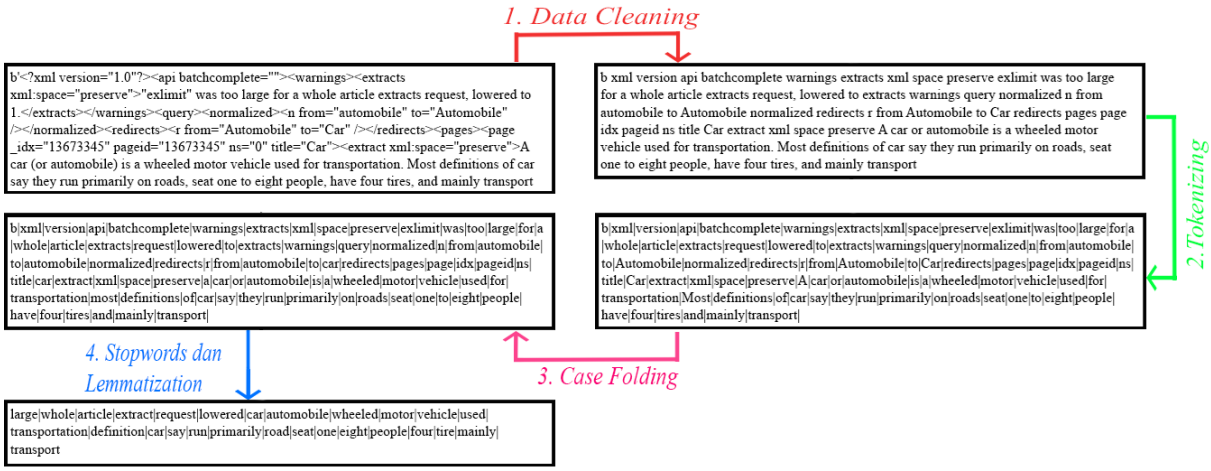
b'<?xml version="1.0"?><api batchcomplete=""><warnings><extracts xml:space="preserve">"exlimit" was too large for a whole article extracts request, lowered to 1.</extracts></warnings><query><normalized><n from="automobile" to="Automobile" /></normalized><redirects><r from="Automobile" to="Car" /></redirects><pages><page _idx="13673345" pageid="13673345" ns="0" title="Car"><extract xml:space="preserve">A car (or automobile) is a wheeled motor vehicle used for transportation. Most definitions of car say they run primarily on roads, seat one to eight people, have four tires, and mainly transport

b'<?xml version="1.0"?><api batchcomplete=""><warnings><extracts xml:space="preserve">"exlimit" was too large for a whole article extracts request, lowered to 1.</extracts></warnings><query><normalized><n from="gem" to="Gem" /></normalized><redirects><r from="Gem" to="Gemstone" /></redirects><pages><page _idx="12806" pageid="12806" ns="0" title="Gemstone"><extract xml:space="preserve">A gemstone (also called a gem, fine gem, jewel, precious stone, or semi-precious stone) is a piece of mineral crystal which, in cut and polished form,

Gambar 3. **Contoh Penggabungan Artikel Wikipedia**

3.3 Preprocessing

Tahap ini akan melakukan *preprocessing*. Pertama teks akan dilakukan proses *Data Cleaning*, yaitu proses untuk menghilangkan karakter-karakter yang bukan huruf. Karakter tersebut dapat berupa tanda baca, angka, operasi, dan aritmatik. Kedua teks akan dilakukan proses *Tokenizing*, yaitu tahap dilakukannya perubahan struktur teks menjadi bentuk *token*. *Token* tersebut dipisahkan berdasarkan tanda spasi dalam teks sehingga dapat terlihat pembagian perkatanya agar dapat dilanjutkan ke proses berikutnya. Ketiga teks akan dilakukan proses *Case Folding*, yaitu proses untuk mengubah bentuk huruf-huruf *uppercase* menjadi *lowercase* sehingga dapat terhindar dari *case sensitive*. Keempat teks akan dilakukan proses *Stopwords* dan *Lemmatization*, yaitu proses untuk menghilangkan kata-kata yang sering muncul atau dianggap tidak memiliki arti dan proses untuk mengembalikan bentuk dasar dari sebuah kata. Untuk lebih jelas gambaran proses *Preprocessing* secara umum dapat dilihat pada Gambar 4.



Gambar 4. **Tahap *Preprocessing***

3.4 Perhitungan PMI*max*

Tahap ini akan melakukan Perhitungan PMI*max*. Pertama melakukan pembentukan matriks untuk membantu proses pencarian nilai dalam proses pembobotan. Dalam proses pembuatan matriks, digunakan kata unik yang diambil dari dataset korpus wikipedia hasil *preprocessing*. Kedua melakukan proses pembobotan. Pembobotan digunakan untuk mendapatkan nilai *co-occurance* atau kemunculan pasangan kata pada saat yang bersamaan pada sebuah konteks. Dalam proses pembobotan menggunakan matriks *term-context* yang telah dibentuk pada proses sebelumnya dan *window size* dengan nilai tertentu. Ketiga akan dilakukan proses perhitungan PMI*max* yang bertujuan untuk mencari nilai *semantic similarity*. Dengan mendapatkan nilai *semantic similarity*, dapat dilihat seberapa erat hubungan dua buah kata. *Input*an sistem adalah berupa korpus wikipedia dan *gold standard*. Setiap pasangan kata akan dilihat nilai kemunculannya untuk mengetahui bobot dari setiap pasangan kata tersebut. Nilai bobot ini berguna untuk memenuhi nilai frekuensi kata 1 dan kata 2 pada persamaan PMI*max*. Selanjutnya sistem akan menghitung total jumlah kata yang ada pada korpus. Dilanjutkan dengan menghitung nilai *sense* pada kata 1 dan kata 2 dengan menggunakan rumus pada persamaan (2). Setelah semua proses dilakukan dan variabel telah terpenuhi, sistem akan langsung menghitung nilai *semantic similarity* menggunakan PMI*max* dengan rumus pada persamaan (1). Dari proses tersebut akan dihasilkan nilai *semantic similarity* antar sepasang kata. Keempat akan dilakukan proses perhitungan korelasi. Pada saat perhitungan korelasi dibutuhkan dua data, yaitu nilai *semantic similarity* pada sistem dan skor *similarity* pada *gold standard*. Kedua data tersebut akan dilakukan perhitungan korelasi menggunakan korelasi *pearson* dengan rumus pada persamaan (3). Nilai korelasi dihitung menggunakan *library* yang sudah ada pada *python* dan akan digunakan untuk mengevaluasi sistem yang dibangun. Untuk contoh perhitungan nilai *sense* untuk kata 1 dan kata 2 dapat dilihat pada Gambar 5. Kemudian untuk contoh perhitungan PMI*max* dapat dilihat pada Gambar 6. Untuk penjelasan mengenai program 1, program 2, program 3, dan program 4 dapat dilihat pada Lampiran 1 sampai dengan Lampiran 4.

Kata pertama "gem" dan kata kedua "jewel" :

- Perhitungan nilai *sense* :

Sense kata pertama : $yw1 = \frac{(\log(fw1)+q)^p}{(\log(700)+q)^p}$

$yw1 = \frac{(\log(40)+10)^{7.5}}{(\log(700)+10)^{7.5}} = \frac{333241624.5}{13841960317} = 0.24$

Sense kata kedua : $yw2 = \frac{(\log(fw2)+q)^p}{(\log(700)+q)^p}$

$yw2 = \frac{(\log(26)+10)^{7.5}}{(\log(700)+10)^{7.5}} = \frac{262185595.2}{1384196032} = 0.18$

Keterangan :
fw1, fw2 = nilai kemunculan kata pertama (gem) dan kata kedua (jewel) dalam korpus.

q = variabel dengan nilai antara range [-6 - 10] berkelipatan 1. Contoh disini menggunakan nilai 10.

p = variabel dengan nilai antara range range [0 - 10] berkelipatan 0,5. Contoh disini menggunakan nilai 7.5

Gambar 5. **Contoh Hitung Nilai *Sense***

Kata pertama "gem" dan kata kedua "jewel" :

- Perhitungan nilai PMI_{max} :

$$PMI_{max}(w_1,w_2) = \log\left(\frac{(fd(w1,w2)-\frac{e^k}{N}(fw1.fw2-\frac{fw1}{yw1}\frac{fw2}{yw2}))N}{\frac{fw1}{yw1}\frac{fw2}{yw2}}\right)$$

$$PMI_{max}(gem,jewel) = \log\left(\frac{(2-\frac{10}{36237}(40.26-\frac{40}{0,24}\frac{26}{0,18}))36287}{\frac{40}{0,24}\frac{26}{0,18}}\right) = 2.54$$

Keterangan :
- fw1, fw2 = nilai kemunculan kata pertama (gem) dan kata kedua (jewel) dalam korpus.
- fd(w1,w2) = nilai frekuensi co-occurrence antara sebuah pasangan kata w1(gem) dan w2(jewel). Pada matriks akhir nilai co-occurence kata gem dan jewel adalah 2.
- w1 = merupakan kata pertama (gem).
- w2 = merupakan kata kedua (jewel).
- e^k = nilai tetapan yaitu 10.
- N = merupakan total jumlah kata pada korpus.
- yw1, yw2 = merupakan nilai sense dari kata pertama (gem) dan kata kedua (jewel).

Gambar 6. **Contoh Hitung Nilai PMI*max***

4. Evaluasi

4.1 Hasil Pengujian

Pengujian dilakukan untuk membandingkan nilai korelasi sistem dari metode yang berbeda yaitu metode PMI dengan metode PMI*max* yang menggunakan korpus bahasa inggris dan tiga *gold standard* yaitu *Miller-Charles, SimLex-999* dan *WordSim-353* sebagai *input* sistem. Setelah korpus dan *gold standard* di*input*kan, maka akan menghasilkan *output* berupa nilai *similarity*. kemudian nilai *similarity* dan *gold standard* tersebut digunakan sebagai *input* sistem untuk mendapatkan nilai korelasi. Perbandingan nilai korelasi yang didapat dari sistem yang dibangun dengan menggunakan PMI dan PMI*max* dapat dilihat pada Tabel 2.

Tabel 2. **Perbandingan nilai korelasi dengan menggunakan PMI dan PMI*max***

Nilai Korelasi (PMI)		
Miller-Charles	**SimLex-999**	**WordSim-353 (Similarity)**
0.42	-0.06	0.16
Nilai Korelasi (PMI*max*)		
Miller-Charles	**SimLex-999**	**WordSim-353 (Similarity)**
0.71	-0.02	0.37

4.2 Analisis Pengaruh Penggunaan *Window Size* terhadap Perhitungan PMI*max*

Jika dilihat pada Tabel 3, terdapat hasil nilai *semantic similarity* untuk beberapa pasangan kata. Untuk mendapatkan nilai *semantic similarity* pada metode ini, diperlukan pencarian nilai bobot (*Co-Occurence*) dengan meng-

gunakan *window size*. Setelah dilakukan pengujian dengan menggunakan 2 nilai *window size* yaitu berukuran 46 dan 88, dihasilkan beberapa peningkatan nilai *similarity* pada beberapa pasangan kata. Peningkatan nilai *similarity*, dipengaruhi oleh kemunculan dari sepasang kata secara bersamaan pada korpus. Saat ukuran *window size* ditingkatkan, maka ada kemungkinan kemunculan sepasang kata secara bersamaan tersebut akan meningkat juga. Contoh pada pasangan kata *"automobile"* dan *"car"* yang ada pada dataset *Miller-Charles*, dengan menggunakan *window size* 46, kemunculan pasangan kata secara bersamaan pada korpus sebanyak 44 kali, sedangkan pada *window size* 88, kemunculan pasangan kata secara bersamaan pada korpus sebanyak 100 kali. Peningkatan kemunculan kata tersebut juga mempengaruhi peningkatan nilai *similarity* pasangan kata tersebut sebesar 0.77. Namun ada beberapa pasangan kata yang nilai *similarity* tetap seperti kata *"brother"* dan *"lad"*, tetapi tidak ada pasangan kata yang nilai *similarity*nya menurun. Untuk grafik semua dataset *gold standard* yang digunakan dapat dilihat pada Lampiran 5 sampai dengan Lampiran 7.

Tabel 3. **Pengaruh Kemunculan Kata Terhadap Nilai *Similarity* Pada Dataset *Miller-Charles***

90 Artikel		
Pasangan Kata	Nilai Bobot (WS46)	Nilai Bobot (WS88)
automobile, car	44	100
asylum, madhouse	0	0
brother, lad	0	0
gem, jewel	2	4
crane, implement	0	0

90 Artikel		
Pasangan Kata	Nilai *Similarity* (WS46)	Nilai *Similarity* (WS88)
automobile, car	4.64	5.41
asylum, madhouse	0	0
brother, lad	2.28	2.28
gem, jewel	3.01	3.43
crane, implement	2.26	2.26

4.3 Analisis Pengaruh Penggunaan Jumlah Artikel terhadap Perhitungan PMI*max dan Pasangan Kata Yang Dapat Dihitung*

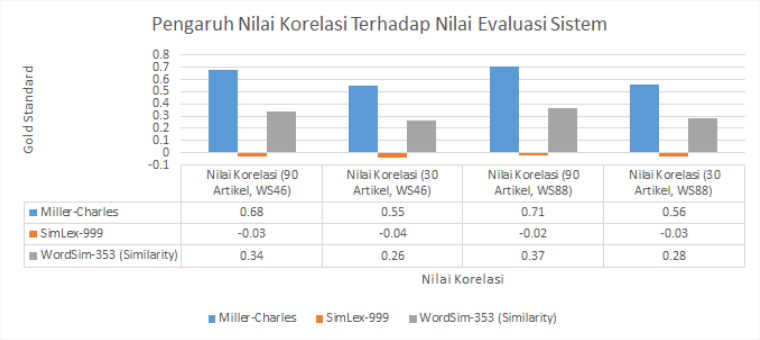
Jika dilihat pada Tabel 4, untuk kata *"automobile"* dan kata *"car"* dengan menggunakan data 90 artikel dengan *window size* berukuran 46 mendapatkan nilai *similarity* sebesar 3.55 dan untuk data 30 artikel dengan *window size* berukuran 46 mendapatkan nilai *similarity* sebesar 3.64. Dapat dilihat nilai *similarity* untuk pasangan kata *"automobile"* dan kata *"car"* lebih besar dengan menggunakan data 30 artikel dibanding dengan menggunakan data 90 artikel dengan selisih 0.09. Oleh karena itu semakin besar jumlah artikel wikipedia yang digunakan belum tentu menghasilkan nilai *similarity* yang lebih tinggi, tergantung dari isi konten artikel tersebut. Pada pengujian dataset *Miller-Charles* dengan jumlah artikel sebanyak 30 judul dapat menghitung pasangan kata sebanyak 14 pasang, dan 90 judul dapat menghitung sejumlah 26 pasangan kata. Untuk melihat pasangan kata yang dapat dihitung dengan menggunakan 30 judul artikel dapat dilihat pada Lampiran 12. Untuk melihat pasangan kata yang dapat dihitung dengan menggunakan 90 judul artikel dapat dilihat pada Lampiran 13.

Tabel 4. **Pengaruh Jumlah Artikel Terhadap Nilai *Similarity***

Pasangan Kata	Nilai *Similarity* (30 Artikel, WS46)	Nilai *Similarity* (90 Artikel, WS46)
automobile, car	3.64	3.55
asylum, madhouse	0	0
brother, lad	2.29	2.28
gem, jewel	2.54	3.01
crane, implement	2.3	2.26

4.4 Analisis Pengaruh Hasil Nilai Korelasi Sebagai Evaluasi Sistem

Dapat dilihat pada Gambar 7, terdapat beberapa nilai korelasi dengan menggunakan 3 jenis *gold standard* yaitu *WordSim-353 (Similarity), SimLex-999,* dan *Miller-Charles*. Korelasi disini menggambarkan seberapa besar akurasi sistem yang dibangun menggunakan metode tertentu. Jika dilihat, nilai korelasi pada *WordSim-353 (Similarity)* dan *Miller-Charles* memiliki nilai positif (lebih besar dari 0). Karena nilai *similarity* pada sistem memiliki arah peningkatan dan penurunan nilai yang searah. Sehingga hal tersebut dapat menghasilkan nilai korelasi positif. Namun nilai korelasi yang dihasilkan belum mencapai 1, hal tersebut disebabkan karena ada beberapa nilai *similarity* dan nilai *gold standard* yang memiliki nilai yang tidak sama. Sedangkan untuk *SimLex-999* memiliki nilai negatif (lebih kecil dari 0). Karena nilai *similarity* pada sistem memiliki arah peningkatan dan penurunan nilai yang tidak searah. Sehingga hal tersebut dapat menghasilkan nilai korelasi negatif. Penyebab adanya nilai korelasi negatif adalah karena masih banyak nilai *similarity* pada sistem yang berbeda jauh dari nilai *gold standard* dan masih banyak juga nilai *similarity* dari pasangan kata yang 0, disebabkan karena salah satu kata pada pasangan kata tersebut tidak ada pada korpus.



Gambar 7. **Grafik Pengaruh Nilai Korelasi Terhadap Evaluasi Sistem**

5. Kesimpulan

Dari pengujian dan analisis yang telah dilakukan pada tugas akhir ini, dapat disimpulkan bahwa :

1. Nilai *similarity* antar kata dipengaruhi oleh kemunculan pasangan kata secara bersamaan.
2. Semakin tinggi ukuran *window size*, maka semakin tinggi peluang nilai *Co-Occurence* (kemunculan pasangan kata secara bersamaan) meningkat.
3. Semakin besar dataset wikipedia yang digunakan maka semakin besar kemungkinan pasangan kata yang dapat dihitung dari dataset *gold standard*.
4. Nilai *similarity* terbaik yang dihasilkan sistem adalah 6.61.
5. Nilai korelasi terbaik yang dihasilkan sistem adalah 0.71 dengan menggunakan *gold standard* dari *Miller-Charles*, *window size* dengan nilai 88, dan jumlah artikel sebanyak 90.

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

1. Menggunakan korpus wikipedia bahasa inggris dengan ukuran file diatas 1MB, karena keterbatasan perangkat yang digunakan pada saat pembangunan sistem sehingga hanya bisa menggunakan file maksimal 1MB.
2. Menggunakan nilai *window size* selain 46 dan 88, karena nilai *similarity* antara pasangan kata yang didapat ketika menggunakan *window size* 46 dan 88 masih ada yang lebih kecil dari *gold standard* sehingga nilai korelasi yang dihasilkan belum mencapai 1.
3. Menggunakan dataset *gold standard* selain *WordSim-353, SimLex-999,* dan *Miller-Charles* dalam membandingkan nilai *similarity*, agar dapat dianalisis lebih jauh dan membuktikan seberapa efektif metode PMI*max* dalam mengukur kesamaan semantik antara pasangan kata.

Daftar Pustaka

[1] Ari Visa. Technology of text mining. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 1–11. Springer, 2001.

[2] Lushan Han, Tim Finin, Paul McNamee, Anupam Joshi, and Yelena Yesha. Improving word similarity by augmenting pmi with estimates of word polysemy. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1307–1322, 2013.

[3] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[4] François Role and Mohamed Nadif. Handling the impact of low frequency events on co-occurrence based measures of word similarity-a case study of pointwise mutual information. In *KDIR*, pages 226–231, 2011.

[5] Mehdi Mohammadi and Nasser GhasemAghaee. Building bilingual parallel corpora based on wikipedia. In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volume 2, pages 264–268. IEEE, 2010.

[6] Thabet Slimani. Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv:1310.8059*, 2013.

[7] Ahmad Pesaranghader, Saravanan Muthaiyah, and Ali Pesaranghader. Improving gloss vector semantic relatedness measure by integrating pointwise mutual information: Optimizing second-order co-occurrence vectors computed from biomedical corpus and umls. In *Informatics and Creative Multimedia (ICICM), 2013 International Conference on*, pages 196–201. IEEE, 2013.

[8] François Gagné. *Descriptive Statistics and Analysis in Biochemical Ecotoxicology*. 2014.

[9] Nian Shong Chok. *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. PhD thesis, University of Pittsburgh, 2010.

[10] Markus Krötzsch, Denny Vr Denny Vrandecic, and Max Völkel. Wikipedia and the semantic web-the missing links. In *Proceedings of Wikimania 2005*. Citeseer, 2005.

[11] Markus Krötzsch, Denny Vrandečić, Max Völkel, Heiko Haller, and Rudi Studer. Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):251–261, 2007.

[12] Ian H Witten and David N Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. 2008.

[13] Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. Testing apsyn against vector cosine on similarity estimation. *arXiv preprint arXiv:1608.07738*, 2016.

[14] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 4(4):665–695, 2015.

[15] S Vijay. Combined method to measure the semantic similarity between words. *International Journal of Soft Computing and Engineering (IJSCE)*, 1:49–54, 2012.