TUGAS WORD2VEC PENGOLAHAN BAHASA ALAMI

INFORMATION RETRIEVAL FOR BPJS SERVICES ARTICLE

KELOMPOK BPJS

Naura Jasmine Azzahra	5026211005
Ramadhanul Husna A. M.	5026211059
Fikri Septa Setiawan	5026211109

KELAS

PENGOLAHAN BAHASA ALAMI (A)

DEPARTEMEN SISTEM INFORMASI

FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SEMESTER GASAL 2024

DAFTAR ISI

Laporan Implementasi Model Word2Vec

Laporan ini menjelaskan detail pelaksanaan Word2Vec untuk analisis teks menggunakan dataset yang diproses dengan pustaka populer seperti NLTK dan Gensim. Model ini dikembangkan untuk mengeksplorasi hubungan semantik antar kata dalam sebuah korpus besar.

1. Pendahuluan

Model Word2Vec adalah metode yang digunakan dalam pembelajaran mesin untuk menghasilkan representasi vektor dari kata-kata, yang memungkinkan pengenalan pola semantik dan hubungan antar kata. Dengan memanfaatkan teknik pembelajaran mendalam, Word2Vec dapat menangkap makna kontekstual dari kata-kata berdasarkan penggunaannya dalam kalimat. Laporan ini berfokus pada implementasi dua metode utama dalam Word2Vec, yaitu CBOW (Continuous Bag of Words) dan Skip-gram, dengan menggunakan data teks yang diambil dari berbagai sumber.

Metode CBOW berfungsi untuk memprediksi kata target berdasarkan kata-kata konteks di sekitarnya, sedangkan Skip-gram melakukan hal sebaliknya, yaitu memprediksi kata-kata konteks berdasarkan kata target. Kedua pendekatan ini memiliki keunggulan masing-masing dan dapat digunakan sesuai dengan kebutuhan analisis yang diinginkan.

Pengaturan lingkungan merupakan langkah penting dalam proses implementasi. Ini melibatkan instalasi pustaka yang diperlukan dan pengunduhan data yang dibutuhkan. Pustaka berikut sangat penting untuk menjalankan analisis Word2Vec:

- NLTK: Digunakan untuk tugas pemrosesan bahasa alami, seperti tokenisasi dan penghilangan stop words.
- Gensim: Digunakan untuk melatih dan bekerja dengan model Word2Vec, menyediakan antarmuka yang efisien untuk mengelola vektor kata.
- Matplotlib dan scikit-learn: Digunakan untuk visualisasi hubungan antar kata dan reduksi dimensi, sehingga hasil analisis dapat dipahami dengan lebih baik.
- Pandas: Digunakan untuk membaca dan mengelola data dalam format yang mudah diakses dan dianalisis.

2. Word2Vec : Contoh Berdasarkan PPT yang Diberikan

Pada tahap awal, percobaan dilakukan menggunakan dataset fetch_20newsgroups, yang merupakan kumpulan data berita yang sering digunakan dalam analisis teks. Dataset ini berisi artikel-artikel dari berbagai kelompok berita, sehingga memberikan konteks yang kaya untuk analisis. Proses pertama yang dilakukan adalah tokenisasi, di mana teks dipecah menjadi kata-kata individual menggunakan pustaka NLTK. Setelah proses tokenisasi, data yang telah diproses kemudian dilatih menggunakan model Word2Vec untuk menghasilkan representasi vektor dari kata-kata tersebut.

Hasil dari percobaan ini diharapkan dapat mencerminkan konsep yang telah diajarkan dalam presentasi sebelumnya. Fine-tuning model dilakukan dengan mencoba beberapa variasi

eksperimen untuk meningkatkan akurasi hasil. Pada percobaan awal, ketika model dilatih dengan data yang tidak diproses secara menyeluruh, terdapat kesalahan dalam menghasilkan analogi. Sebagai contoh, ketika diuji dengan analogi ('man', 'woman', 'king'), model memprediksi kata 'weaver', yang jelas tidak sesuai dengan konteks.

Kesalahan ini dapat dijelaskan oleh dua faktor utama: keterbatasan jumlah data dan pelatihan model dari awal tanpa referensi yang cukup. Untuk memperbaiki hasil ini, langkah selanjutnya adalah mencari dataset yang lebih sesuai dan kaya informasi. Dalam hal ini, dataset Text8 digunakan sebagai alternatif. Setelah model dimuat dengan dataset baru ini dan menjalani pelatihan ulang, hasil dari analogi yang sama menunjukkan perbaikan signifikan; model kini memprediksi kata 'empress' untuk analogi ('man', 'woman', 'king'), yang lebih tepat dan relevan.

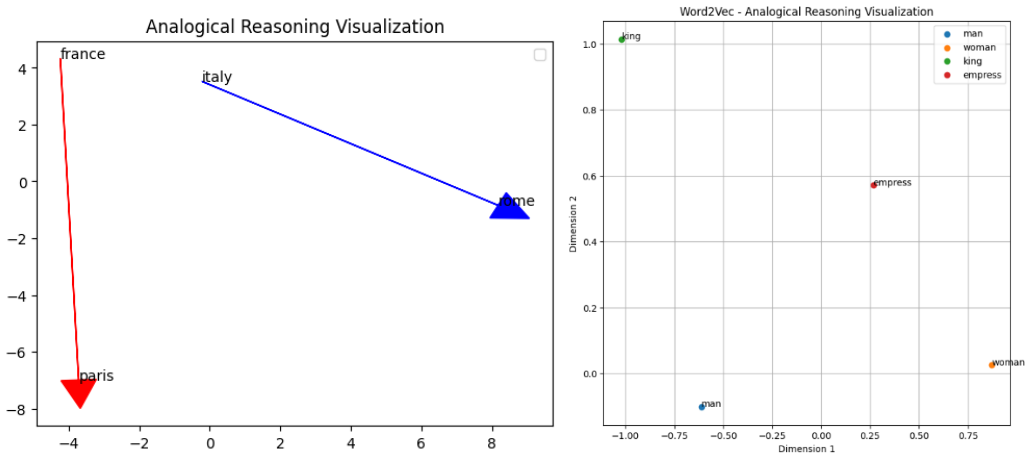
Output dari kedua percobaan tersebut menunjukkan perbedaan yang mencolok dalam kemampuan model dalam memahami hubungan semantik antar kata. Pada percobaan pertama, output untuk analogi adalah:

```
The predicted word for the analogy ('man', 'woman', 'king') is: weaver
```

Sedangkan setelah menggunakan dataset Text8, outputnya adalah:

```
The predicted word for the analogy ('man', 'woman', 'king') is: empress
```

Perbaikan ini menunjukkan bahwa kualitas dan kuantitas data pelatihan sangat mempengaruhi kemampuan model dalam memahami dan menghasilkan hubungan semantik antar kata. Dengan menggunakan dataset yang lebih besar dan terstruktur dengan baik, model dapat belajar dari lebih banyak konteks, sehingga meningkatkan akurasi prediksi dalam tugas-tugas analisis teks.Hasilnya menunjukkan analogi yang sesuai. Model yang digunakan telah dilatih menggunakan data yang banyak sehingga hasil yang didapatkan tentunya akan lebih baik. Dalam beberapa hal percobaan lainnya kita juga dapat melihat vektor dari analogi yang telah dilakukan dan hasilnya sebagai berikut :



Selain itu, dalam beberapa percobaan lainnya, kita juga dapat melihat representasi vektor dari analogi yang telah dilakukan. Misalnya, ketika mencari kata-kata yang paling mirip dengan 'news', hasilnya menunjukkan kata-kata seperti:

```
Words most similar to 'news': [('articles', 0.7407447099685669), ('usenet', 0.714850962162017
8), ('paper', 0.7032508850097656), ('previous', 0.7015998363494873), ('reports', 0.700215160846
7102), ('reviews', 0.6989465951919556), ('group', 0.6947528123855591), ('topic', 0.692541420459
7473), ('recent', 0.6910982131958008), ('list', 0.6882311105728149)]
```

3. TUGAS Pre Tained Word2Vec Indonesian Model

Percobaan ini berfokus pada penggunaan model Word2Vec yang telah dilatih sebelumnya, yaitu idwiki_word2vec_300, untuk menganalisis kata-kata yang berkaitan dengan BPJS. Model ini diharapkan dapat memberikan hasil yang lebih relevan dan akurat dalam konteks bahasa Indonesia, terutama terkait dengan istilah-istilah dalam bidang kesehatan dan jaminan sosial.

A. Words Similarity (Sinonim Kata)

Pertama, kami melakukan analisis sinonim untuk beberapa kata kunci yang berkaitan dengan BPJS. Dengan menggunakan model Word2Vec, kami mencari kata-kata yang memiliki kemiripan semantik dengan kata-kata berikut:

- 'sehat': Hasilnya adalah ['bugar', 'mandiri', 'harmonis', 'bersih', 'kesegaran'].
- 'bpjs': Hasilnya adalah ['jamsostek', 'askes', 'jamkesmas', 'ketenagakerjaan', 'tabungan'].
- 'layan': Hasilnya adalah ['ayuhai', 'praktak', 'kesangkut', 'crosthwaite', 'nawangsari'].
- 'program': Hasilnya adalah ['programnya', 'progam', 'comdev', 'acara', 'paket'].
- 'sakit': Hasilnya adalah ['rs', 'rsu', 'nyeri', 'bersalin', 'rsud'].
- 'perintah': Hasilnya adalah ['perintahnya', 'titah', 'instruksi', 'anjuran', 'nasihat'].
- 'data': Hasilnya adalah ['informasi', 'datanya', 'metadata', 'pendataan', 'basisdata'].
- 'jamin': Hasilnya adalah ['djamin', 'brigjend', 'sadeli', 'dasuki', 'ginting'].

Hasil analisis menunjukkan bahwa beberapa sinonim seperti "bugar" untuk "sehat" dan "jamsostek" untuk "bpjs" cukup relevan. Namun, terdapat beberapa kata yang tidak masuk akal dalam konteks BPJS, seperti "ayuhai" dan "gorat". Hal ini menunjukkan bahwa model mungkin belum memiliki data yang cukup untuk menangkap makna yang tepat terkait istilah-istilah spesifik dalam konteks BPJS.

```
Similarity between 'sehat' and 'bpjs': 0.34401753544807434
Similarity between 'sehat' and 'layan': 0.05648801848292351
Similarity between 'sehat' and 'program': 0.22241538763046265
Similarity between 'sehat' and 'sakit': 0.18878547847270966
Similarity between 'sehat' and 'perintah': 0.02385062538087368
Similarity between 'sehat' and 'data': 0.15633940696716309
Similarity between 'sehat' and 'jamin': 0.1041366308927536
Similarity between 'bpjs' and 'layan': 0.1459125131368637
Similarity between 'bpjs' and 'program': 0.28401386737823486
Similarity between 'bpjs' and 'sakit': 0.17045281827449799
Similarity between 'bpjs' and 'perintah': 0.040987513959407806
Similarity between 'bpjs' and 'data': 0.16501864790916443
Similarity between 'bpjs' and 'jamin': 0.11866355687379837
Similarity between 'layan' and 'program': -0.013493876904249191
Similarity between 'layan' and 'sakit': 0.0206647627055645
Similarity between 'layan' and 'perintah': -0.03285708278417587
Similarity between 'layan' and 'data': 0.034656040370464325
Similarity between 'layan' and 'jamin': 0.1350693702697754
Similarity between 'program' and 'sakit': 0.09969556331634521
Similarity between 'program' and 'perintah': 0.1917981505393982
Similarity between 'program' and 'data': 0.26221764087677
Similarity between 'program' and 'jamin': -0.036139026284217834
Similarity between 'sakit' and 'perintah': 0.10108301788568497
Similarity between 'sakit' and 'data': 0.013396796770393848
Similarity between 'sakit' and 'jamin': 0.06089222431182861
Similarity between 'perintah' and 'data': 0.2482822984457016
Similarity between 'perintah' and 'jamin': -0.02393687702715397
Similarity between 'data' and 'jamin': -0.06977172195911407
```

Dalam percobaan ini, kami mengevaluasi kesamaan antar kata yang berkaitan dengan BPJS menggunakan model Word2Vec yang telah dilatih sebelumnya. Proses ini melibatkan pemuatan model dan perhitungan skor kesamaan untuk pasangan kata yang telah ditentukan, seperti 'sehat', 'bpjs', 'layan', dan lainnya. Hasil dari evaluasi menunjukkan variasi dalam tingkat kesamaan antar kata, di mana beberapa pasangan menghasilkan skor yang lebih tinggi dibandingkan yang lain. Misalnya, kesamaan antara 'sehat' dan 'bpjs' adalah 0.344, menunjukkan adanya hubungan semantik yang moderat antara kedua kata tersebut. Namun, beberapa pasangan seperti 'layan' dan 'program' menunjukkan skor negatif (-0.013), yang mengindikasikan bahwa kata-kata tersebut tidak memiliki hubungan semantik yang jelas dalam konteks model.

Hasil ini memberikan informasi tentang bagaimana model menangkap makna kata dalam konteks bahasa Indonesia dan menunjukkan bahwa meskipun model dapat memberikan informasi berguna, ada juga batasan dalam hal data yang digunakan untuk pelatihan. Beberapa hasil sinonim dan analogi yang dihasilkan oleh model menunjukkan bahwa ada kata-kata yang tidak relevan atau tidak masuk akal dalam konteks BPJS, seperti 'gorat' untuk analogi ('program', 'data', 'jamin'), yang menandakan bahwa model mungkin belum sepenuhnya memahami hubungan semantik dalam konteks spesifik ini.

B. Fill in The Blank (Pengisian Kalimat)

Selanjutnya, kami melakukan percobaan pengisian kalimat (fill in the blank). Berikut adalah contoh kalimat yang diuji:

a. Kalimat Asli: Apakah BPJS dapat ___ semua penyakit yang diderita oleh warga

   Kalimat Terisi: Apakah BPJS dapat "**jika**" semua penyakit yang diderita oleh warga

b. Kalimat Asli: Penyakit Demam Berdarah telah ___ di seluruh pelosok kota Indonesia

Kalimat Terisi: Penyakit Demam Berdarah telah "**wabah**" di seluruh pelosok kota Indonesia

Hasil dari percobaan pertama menunjukkan bahwa pengisian kalimat tidak tepat, sedangkan pada percobaan kedua hasilnya masih dapat diterima karena kata "wabah" berhubungan dengan penyebaran penyakit.

C. Predicted Word for Analogy (Analogi)

Percobaan selanjutnya adalah analisis analogi. Kami menguji beberapa analogi menggunakan model Word2Vec:

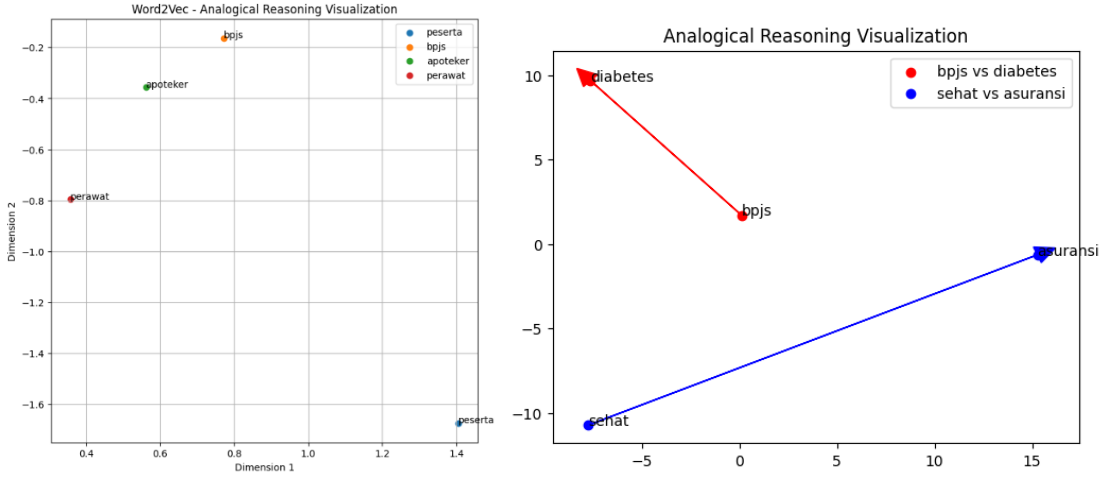
- Analogi 1: ('sehat', 'sakit', 'bpjs') menghasilkan prediksi kata: **rs**
- Analogi 2: ('program', 'data', 'jamin') menghasilkan prediksi kata: **gorat**
- Analogi 3: ('layan', 'bpjs', 'perintah') menghasilkan prediksi kata: **rekomendasi**

Model berhasil memetakan beberapa analogi dengan baik, seperti pada analogi pertama, di mana rs (rumah sakit) berhubungan langsung dengan BPJS. Namun, hasil dari analogi kedua (gorat) tidak relevan dalam konteks jaminan. Pada analogi ketiga, hasil rekomendasi cukup longgar tetapi masih bisa diterima sebagai instruksi terkait layanan.

D. Visualisasi

Untuk lebih memahami hubungan antar kata, kami juga melakukan visualisasi dari hasil analogi menggunakan teknik reduksi dimensi seperti PCA atau t-SNE. Visualisasi ini membantu dalam melihat bagaimana kata-kata tersebut terdistribusi dalam ruang vektor dan memberikan gambaran mengenai hubungan semantik antar kata.

Secara keseluruhan, meskipun model Word2Vec ini menunjukkan ketepatan di beberapa kasus, terdapat juga kesalahan dalam beberapa prediksi. Hal ini menyoroti pentingnya pemilihan dataset yang tepat dan pelatihan model dengan data yang cukup untuk meningkatkan akurasi dalam aplikasi nyata di bidang analisis teks dan pemrosesan bahasa alami. :



4. Update Word2Vec Model with BPJS Data Non Clean

Percobaan ini berfokus pada pembaruan model Word2Vec yang telah dilatih sebelumnya, yaitu idwiki_word2vec_300, dengan menggunakan data berita mengenai BPJS yang belum dibersihkan. Tujuan dari pembaruan ini adalah untuk meningkatkan pemahaman model terhadap istilah-istilah yang berhubungan dengan BPJS, mengingat hasil percobaan sebelumnya menunjukkan bahwa model tersebut kekurangan informasi terkait topik ini.

A. Pembaruan Model dengan Data BPJS Non Clean

Langkah awal dalam proses ini adalah memuat model Word2Vec yang sudah ada dan mempersiapkan data berita mengenai BPJS. Data tersebut akan ditokenisasi untuk memecah teks menjadi kata-kata individual sebelum ditambahkan ke dalam model. Proses ini penting agar model dapat belajar dari konteks yang lebih kaya dan relevan.

Setelah mempersiapkan data, kami melakukan pemuatan model dan pembaruan vocabulary serta pelatihan model dengan data baru. Berikut adalah langkah-langkah yang diambil:

```
import pandas as pd
# Load the pre-trained Word2Vec model
model = Word2Vec.load("idwiki_word2vec_300.model")

# Load the Excel file
data_path = "/content/data_ready_final_v2.xlsx"
df = pd.read_excel(data_path)

# Tokenize the text in the 'text_berita' column and convert to lowercase
df['tokenized_text'] = df['text_berita'].apply(lambda x: word_tokenize(x.lower()))

# Prepare the tokenized text for model update and training
new_data = df['tokenized_text'].tolist()

# Update vocabulary and train the model
model.build_vocab(new_data, update=True)
model.train(new_data, total_examples=len(new_data), epochs=10)
```

B. Words Similarity (Sinonim Kata)

Setelah pembaruan model, kami melakukan analisis sinonim untuk beberapa kata kunci yang berkaitan dengan BPJS. Dengan menggunakan model Word2Vec, kami mencari kata-kata yang memiliki kemiripan semantik dengan kata-kata berikut:

- 'sehat': Hasilnya adalah ['bugar', 'kesegaran', 'walafiat', 'kondusif', 'harmonis'].
- 'bpjs': Hasilnya adalah ['kesehatan', '.', 'yang', 'ini', 'jkn'].
- 'layan': Hasilnya adalah ['ayuhai', 'praktak', 'kesangkut', 'crosthwaite', 'nawangsari'].

- 'program': Hasilnya adalah ['programnya', 'jaminan', 'progam', 'program-program', 'kegiatan'].
- 'sakit': Hasilnya adalah ['rs', 'sakit.advertisementscroll', 'sakitnya', 'mulas', 'pasien'].
- 'perintah': Hasilnya adalah ['perintahnya', 'titah', 'instruksi', 'anjuran', 'nasihat'].
- 'data': Hasilnya adalah ['datanya', 'data-data', 'pendataan', 'informasi', 'monografi'].
- 'jamin': Hasilnya adalah ['djamin', 'peranginangin', 'japat', 'brigjend', 'ginting'].

Hasil analisis menunjukkan bahwa beberapa sinonim seperti "bugar" untuk "sehat" dan "kesehatan" untuk "bpjs" cukup relevan. Namun, terdapat beberapa kata yang tidak masuk akal dalam konteks BPJS, seperti "ayuhai" dan "gorat". Hal ini menunjukkan bahwa model mungkin belum memiliki data yang cukup untuk menangkap makna yang tepat terkait istilah-istilah spesifik dalam konteks BPJS.

C. Evaluate Similarity Words

Dalam analisis ini, kami mencari kata-kata yang memiliki kemiripan semantik dengan beberapa kata kunci penting terkait BPJS. Berikut adalah hasil dari pencarian sinonim untuk kata-kata tersebut:

- Similar words to 'sehat':
  - Hasilnya adalah [('bugar', 0.5169265270233154), ('kesegaran', 0.4192076623439789), ('walafiat', 0.4129122495651245), ('kondusif', 0.40798306465148926), ('harmonis', 0.4057063162326813)].
- Similar words to 'dokter':
  - Hasilnya adalah [('spesialis', 0.549913763999939), ('psikiater', 0.5427910685539246), ('dokternya', 0.5196490287780762), ('perawat', 0.5195528864860535), ('psikolog', 0.5127800703048706)].
- Similar words to 'bpjs':
  - Hasilnya adalah [('kesehatan', 0.6084312796592712), ('.', 0.4963931739330292), ('yang', 0.4959421753883362), ('ini', 0.4594666361808777), ('jkn', 0.45702192187309265)].

Hasil analisis kata-kata yang serupa menunjukkan adanya asosiasi semantik antar kata dalam konteks kesehatan. Kata "sehat" memiliki kemiripan dengan "bugar" yang mengarah pada kondisi fisik yang baik, dan "kesegaran" serta "walafiat" yang berkaitan dengan kesejahteraan tubuh. Kata "dokter" dihubungkan dengan profesi terkait seperti "spesialis", "psikiater", dan "perawat", serta variasi dalam penyebutan profesi tersebut seperti "dokternya" dan "psikolog". Sementara itu, "bpjs" terkait erat dengan istilah kesehatan dan sistem layanan seperti "jkn".

Hasil ini menunjukkan bahwa model telah mendapatkan informasi yang lebih luas dibandingkan sebelumnya, sehingga hasil yang ditunjukkan cukup baik dalam konteks BPJS dan kesehatan.

D. Fill in The Blank (Pengisian Kalimat)

Selanjutnya, kami melakukan percobaan pengisian kalimat (fill in the blank). Berikut adalah contoh kalimat yang diuji:

a. Kalimat Asli: Apakah BPJS dapat ___ semua penyakit yang diderita oleh warga?

   Kalimat Terisi: Apakah BPJS dapat "jika" semua penyakit yang diderita oleh warga.

b. Kalimat Asli: Penyakit Demam Berdarah telah ___ di seluruh pelosok kota Indonesia.

   Kalimat Terisi: Penyakit Demam Berdarah telah "malaria" di seluruh pelosok kota Indonesia.

Hasil dari percobaan pertama menunjukkan bahwa pengisian kalimat tidak tepat, sedangkan pada percobaan kedua hasilnya masih dapat diterima karena kata "malaria" berhubungan dengan penyebaran penyakit.Pada kalimat pertama, "penanganan" menggantikan ruang kosong dengan makna yang sesuai, karena "penanganan biaya pengobatan" menggambarkan tindakan atau proses yang berkaitan dengan biaya medis. Namun, pada kalimat kedua, penggunaan "malaria" tidak tepat karena "malaria" adalah nama penyakit. Hal ini menunjukkan bahwa model masih kurang baik dalam mengisi blank meskipun pada kaliamt perta menunjukkan hasil yang baik

E. Analogycal (Analogi)

Selanjutnya, dilakukan percobaan analisis analogi menggunakan model Word2Vec. Berikut adalah hasil dari beberapa percobaan yang dilakukan:

- Analogi 1: ('sehat', 'sakit', 'bpjs') menghasilkan prediksi kata: rs.
- Analogi 2: ('program', 'data', 'jamin') menghasilkan prediksi kata: basarah.
- Analogi 3: ('bpjs', 'sehat', 'sakit') menghasilkan prediksi kata: sakit.advertisementscroll.

Pada percobaan ini, hasil yang diperoleh menunjukkan bahwa model mengalami kesulitan dalam memberikan prediksi yang tepat. Hasil dari analogi kedua dan ketiga, khususnya, sangat melenceng dari konteks yang diharapkan. Hal ini disebabkan oleh kurangnya praproses data pada dataset yang ditambahkan, sehingga muncul kata-kata yang seharusnya tidak relevan atau penting.

Percobaan berikutnya:

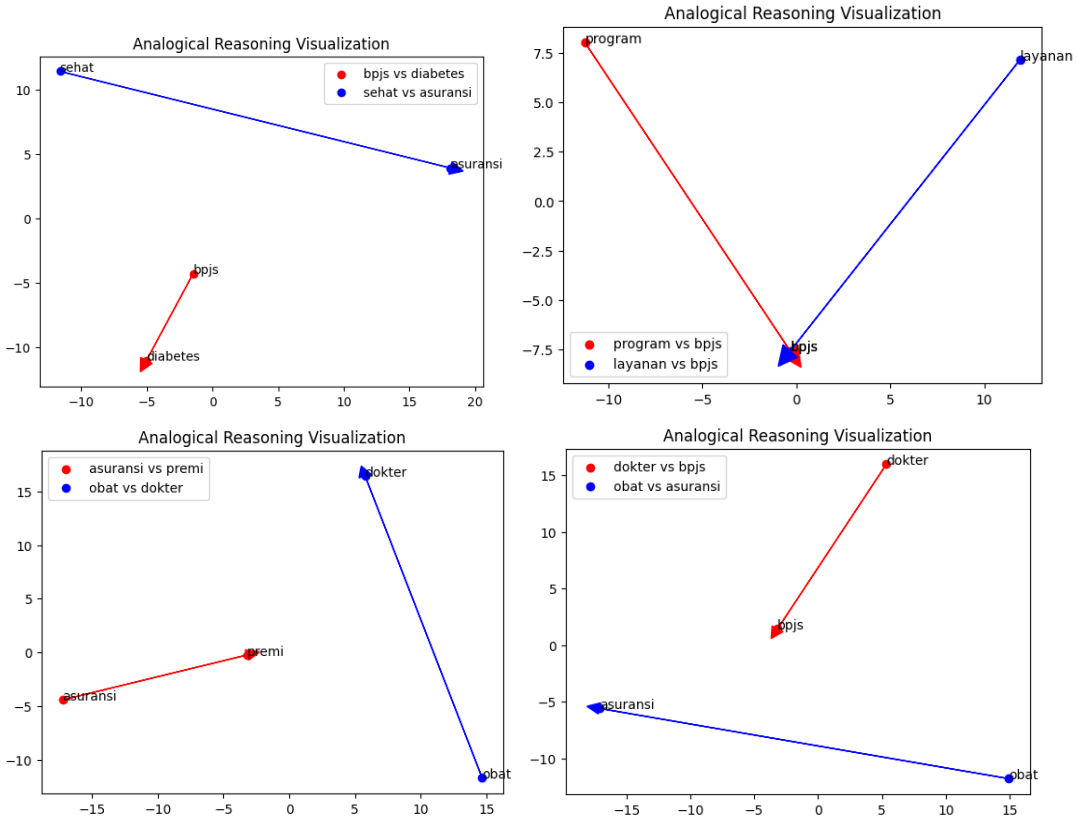
- Analogi 4: ('sehat', 'sakit', 'manfaat') menghasilkan prediksi kata: rs.
- Analogi 5: ('program', 'layanan', 'jaminan') menghasilkan prediksi kata: pelayanan.

- Analogi 6: ('dokter', 'pengobatan', 'bpjs') menghasilkan prediksi kata: kesehatan.
- Analogi 7: ('asuransi', 'bpjs', 'perawatan') menghasilkan prediksi kata: kesehatan.
- Analogi 8: ('peserta', 'bpjs', 'fasilitas') menghasilkan prediksi kata: fktp.
- Analogi 9: ('kesehatan', 'pemerintah', 'jaminan') menghasilkan prediksi kata: perlindungan.

Hasil analisis analogi ini menunjukkan hubungan semantik dalam konteks kesehatan. Misalnya, "rs" (rumah sakit) terkait dengan "sehat" dan "sakit", sedangkan "pelayanan" menggambarkan hubungan antara "program" dan "jaminan". Kata "kesehatan" muncul dalam beberapa analogi, termasuk dengan "dokter", "bpjs", dan "asuransi". Kata "fktp" diprediksi sebagai fasilitas untuk peserta BPJS, sementara "perlindungan" berkaitan dengan jaminan kesehatan dari pemerintah.

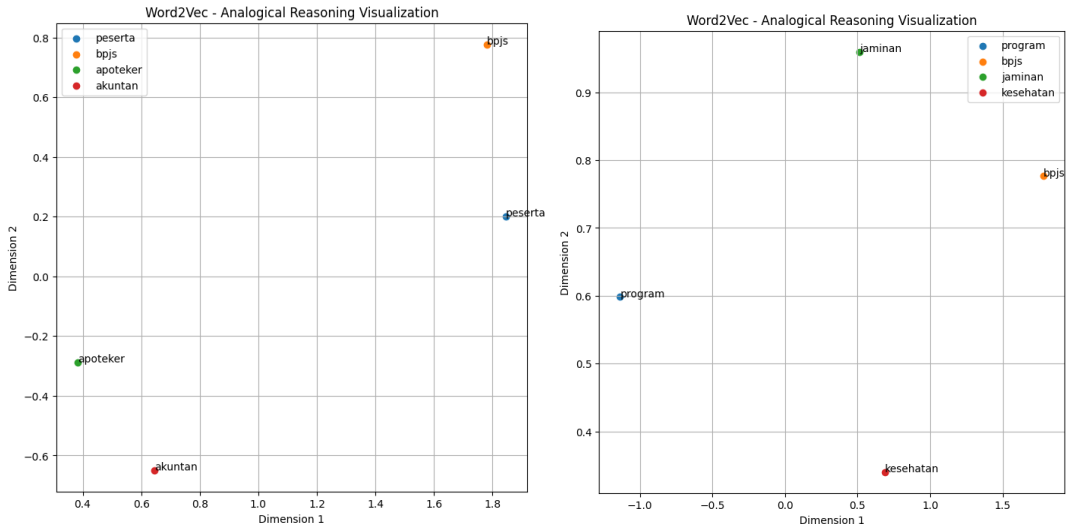
Pada percobaan kedua ini, hasil yang ditunjukkan cukup baik karena jika dianalogikan masih masuk akal, meskipun perlu mengambil cakupan yang lebih luas atau lebih general untuk meningkatkan relevansi dan akurasi hasil.

Selanjutnya yaitu beberapa visualisasi hasil dari analogical :



Secara keseluruhan, visualisasi analogical reasoning ini menunjukkan bahwa model Word2Vec berhasil menangkap hubungan semantik antar kata dalam konteks kesehatan, jaminan sosial, dan layanan asuransi. Hubungan-hubungan seperti "bpjs-diabetes", "program-bpjs", "dokter-bpjs", serta "asuransi-premi" semuanya sangat relevan dengan topik yang dianalisis. Meskipun ada beberapa hasil yang mungkin bisa diperbaiki lebih lanjut dengan data tambahan atau praproses data lebih baik, hasil

visualisasi ini secara umum cukup memadai untuk menggambarkan bagaimana model memahami konsep-konsep dalam domain kesehatan dan jaminan sosial.



5. Update Word2Vec Model with BPJS Data Clean

Setelah melakukan percobaan dengan data yang tidak dibersihkan (non-clean), sekarang dilakukan percobaan dengan data yang telah dipraproses (clean). Langkah-langkah yang diambil serupa dengan sebelumnya, yaitu memperbarui model idwiki_word2vec_300.model menggunakan data yang telah dipilih dan dibersihkan. Pembaruan model dilakukan dari awal, sehingga data non-clean tidak termasuk dalam model yang diperbarui.

Langkah-Langkah:

- Pemuatan Model: Model idwiki_word2vec_300 yang telah dilatih sebelumnya dimuat.
- Pemuatan Data: Data berita mengenai BPJS yang telah dibersihkan dimuat dari file Excel.
- Tokenisasi: Teks dalam kolom text_berita_clean ditokenisasi dan diubah menjadi huruf kecil.
- Pembaruan Vocabulary dan Pelatihan Ulang: Model diperbarui dengan vocabulary baru dari data BPJS yang bersih, dan dilatih ulang selama 10 epoch.

A. Words Sinonymitas

Setelah pembaruan model, dilakukan percobaan untuk mencari sinonim dari beberapa kata kunci terkait BPJS. Berikut adalah hasilnya:

- Sinonim untuk 'sehat': ['bpjs', 'layan', 'jkn', 'jamin', 'pungkas']
- Sinonim untuk 'bpjs': ['sehat', 'layan', 'jamin', 'kait', 'jkn']
- Sinonim untuk 'layan': ['sehat', 'bpjs', 'jkn', 'jamin', 'jknkis']
- Sinonim untuk 'program': ['progam', 'jamin', 'programprogram', 'programnya', 'cakup']
- Sinonim untuk 'sakit': ['rs', 'nyeri', 'rsu', 'rsud', 'bersalin']
- Sinonim untuk 'perintah': ['perintahnya', 'titah', 'anjuran', 'instruksi', 'pemerintah']
- Sinonim untuk 'data': ['datanya', 'informasi', 'analisa', 'pendataan', 'validitas']
- Sinonim untuk 'jamin': ['sehat', 'selenggara', 'bpjs', 'jkn', 'layan']

Hasil analisis sinonim menunjukkan bahwa kata-kata yang terhubung memiliki makna yang serupa dalam konteks tertentu:

- Kata "sehat" sering berasosiasi dengan "bpjs" dan "layan", yang masuk akal dalam konteks layanan kesehatan.
- Kata "bpjs" terkait erat dengan istilah seperti "sehat", "layan", dan "jamin", mencerminkan hubungan antara BPJS dan jaminan kesehatan.
- Kata "layan" juga memiliki hubungan kuat dengan "bpjs" dan "jkn", yang relevan dalam konteks penyediaan layanan kesehatan.
- Kata "sakit" sering diasosiasikan dengan rumah sakit (rs) dan gejala seperti nyeri, menunjukkan hubungan semantik yang logis.
- Kata "perintah" berhubungan dengan istilah seperti "titah" dan "instruksi", mencerminkan perintah dari otoritas atau pemerintah.
- Kata "data" terkait erat dengan istilah seperti "informasi" dan "analisa", menunjukkan pengumpulan dan pemrosesan informasi.
- Kata "jamin", seperti yang diharapkan, terkait erat dengan istilah-istilah seperti "sehat", "bpjs", dan "jkn", mencerminkan jaminan layanan kesehatan.

Hasil analisis sinonim menunjukkan bahwa kata-kata yang terhubung memiliki makna yang serupa dalam konteks tertentu. Misalnya, "sehat" dan "bpjs" sering berasosiasi dalam konteks layanan kesehatan, sementara "layan" terkait dengan "bpjs", "jkn", dan "jamin" dalam konteks penyediaan layanan. "Sakit" sering diasosiasikan dengan "rs" dan "nyeri", sedangkan "perintah" berkaitan dengan "titah" dan "instruksi", mencerminkan perintah dari otoritas. Kata "data" berhubungan dengan "informasi" dan "analisa", yang menunjukkan pengumpulan dan pemrosesan informasi, sedangkan

"jamin" terkait dengan kata-kata seperti "sehat", "bpjs", dan "jkn", yang mencerminkan jaminan layanan kesehatan. Hasil ini menunjukkan bahwa model menangkap informasi yang lebih banyak terkait BPJS namun disini beberapa makan tidak sesuai dengan sinonim yang diminta. Hal ini tentunya juga akan berkaitan dengan parameter disaat melakukan update.

B. Evaluate Similarity Words

Selanjutnya, dilakukan percobaan untuk mengevaluasi kesamaan kata (similar words) menggunakan model Word2Vec yang telah diperbarui dengan data BPJS yang bersih. Berikut adalah hasil dari pencarian kata-kata yang serupa:

- Similar words to 'sehat':
    - [('bpjs', 0.7059), ('layan', 0.6425), ('jkn', 0.6201), ('jamin', 0.5771), ('pungkas', 0.5148)]
- Similar words to 'dokter':
    - [('apoteker', 0.5354), ('perawat', 0.5299), ('ginekolog', 0.5140), ('bidan', 0.5106), ('paramedis', 0.5063)]
- Similar words to 'bpjs':
    - [('sehat', 0.7059), ('layan', 0.6121), ('jamin', 0.5648), ('kait', 0.5626), ('jkn', 0.5289)]

Hasil ini menunjukkan bahwa kata "sehat" berkaitan erat dengan istilah seperti "bpjs", "layan", dan "jkn", yang semuanya relevan dalam konteks layanan kesehatan di Indonesia. Demikian pula, kata "dokter" memiliki hubungan semantik yang kuat dengan profesi medis lainnya seperti "apoteker", "perawat", dan "ginekolog". Sementara itu, kata "bpjs" juga berhubungan erat dengan istilah-istilah terkait jaminan kesehatan seperti "sehat" dan "jamin". Secara keseluruhan, model ini menunjukkan peningkatan dalam menangkap hubungan semantik dibandingkan dengan model sebelumnya.

C. Fill in The Blank (Pengisian Kalimat)

Percobaan berikutnya adalah pengisian kalimat (fill in the blank). Berikut adalah contoh kalimat yang diuji:

- Kalimat Asli: BPJS Kesehatan memberikan perlindungan kepada peserta untuk ___ biaya pengobatan.

  Kalimat Terisi: BPJS Kesehatan memberikan perlindungan kepada peserta untuk penanganan biaya pengobatan.

- Kalimat Asli: Penyakit Demam Berdarah telah ___ di seluruh pelosok kota Indonesia.

Kalimat Terisi: Penyakit Demam Berdarah telah wabah di seluruh pelosok kota Indonesia.

Dari percobaan yang dilakukan model mampu mengisi blank dengan baik dan kaliamt tersebut bisa memiliki makna yang tepat. Dimana adanya penanganan biaya pengobatan dan wabah di seluruh pelosok kota Indonesia. Hal ini menunjukkan adanya peningkatan pada model dalam hal menangkap inforamsi terkait BPJS

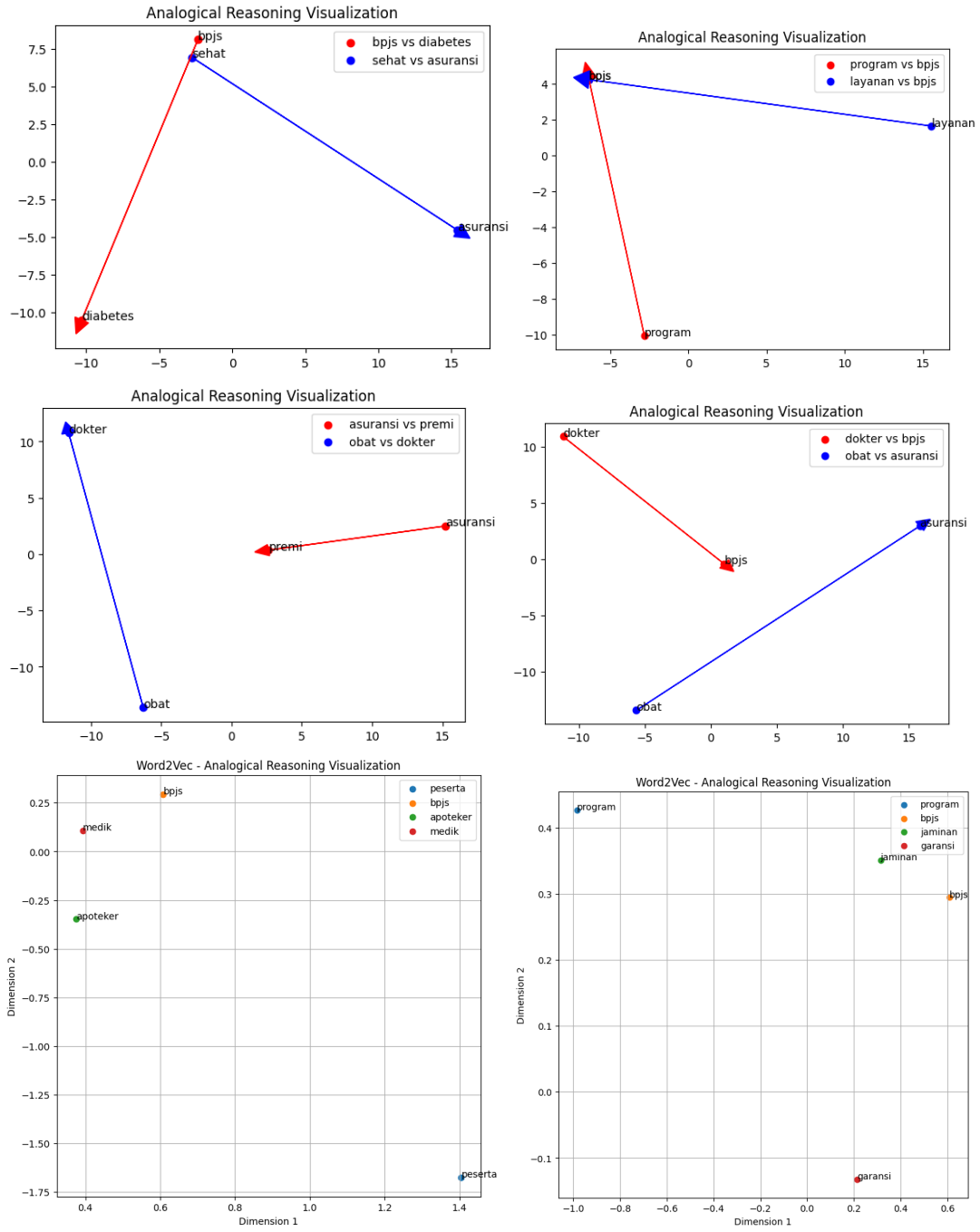
C. Analogycal Words

Percobaan terakhir adalah analisis analogi menggunakan beberapa pasangan kata untuk melihat bagaimana model menangkap hubungan semantik antar kata.

- The predicted word for the analogy ('sehat', 'sakit', 'bpjs') is: rs.
- The predicted word for the analogy ('program', 'data', 'jamin') is: bpjs.
- The predicted word for the analogy ('bpjs', 'sehat', 'sakit') is: nyeri.
- The predicted word for the analogy ('sehat', 'sakit', 'manfaat') is: komplikasi.
- The predicted word for the analogy ('program', 'layanan', 'jaminan') is: garansi.
- The predicted word for the analogy ('dokter', 'pengobatan', 'bpjs') is: sehat.
- The predicted word for the analogy ('asuransi', 'bpjs', 'perawatan') is: pemeriksaan.
- The predicted word for the analogy ('peserta', 'bpjs', 'fasilitas') is: layan.
- The predicted word for the analogy ('kesehatan', 'pemerintah', 'jaminan') is: kompensasi.

Hasil analisis analogi ini menunjukkan bahwa model sudah dapat mengidentifikasi hubungan yang baik antar kata dalam konteks BPJS dan layanan terkait. Prediksi kata-kata seperti "rs", "nyeri", "komplikasi", "garansi", dan "pemeriksaan" menunjukkan kemampuan model dalam memahami istilah-istilah yang saling berhubungan. Selain itu, hubungan antara kata-kata seperti "bpjs", "sehat", dan "sakit" berhasil diprediksi dengan akurat, yang mengindikasikan bahwa model sudah cukup baik dalam menggambarkan konsep-konsep terkait BPJS. Secara keseluruhan, hasil analisis ini menunjukkan bahwa model mampu menghasilkan prediksi yang relevan dan sesuai dengan konteks yang diberikan.

Selanjutnya yaitu beberapa visualisasi hasil dari analogical :



Secara keseluruhan, visualisasi analogical reasoning ini menunjukkan bahwa model Word2Vec mampu menangkap hubungan semantik antar kata-kata yang relevan dalam konteks layanan kesehatan, jaminan sosial, serta profesi medis. Hubungan-hubungan seperti "bpjs-diabetes", "program-bpjs", "dokter-bpjs", serta "asuransi-premi" semuanya sangat relevan dengan topik yang dianalisis. Model berhasil memahami konsep-konsep penting dalam domain kesehatan, meskipun masih ada ruang untuk peningkatan lebih lanjut melalui fine-tuning atau penggunaan dataset yang lebih besar untuk meningkatkan akurasi prediksi.

Kesimpulan

Berdasarkan percobaan yang telah dilakukan, Word2Vec memerlukan data yang banyak dan berkualitas untuk menghasilkan model yang baik. Sebanyak apapun data yang dimiliki, jika data tersebut tidak mengandung konteks yang relevan, maka hasilnya tidak akan sesuai dengan yang diinginkan. Hasil percobaan menunjukkan bahwa model yang dilatih dari awal menghasilkan performa yang lebih buruk dibandingkan model yang telah dilatih sebelumnya, seperti model yang menggunakan data text8. Hal ini juga berlaku pada model idwiki_word2vec_300.model. Meskipun model ini sudah sangat baik karena dilatih menggunakan data Wikipedia, hasilnya tetap tidak optimal jika konteks yang dibutuhkan tidak ada dalam data pelatihan.

Setelah melakukan pembaruan data dengan data BPJS yang telah di-scraping sebelumnya, hasil yang dikeluarkan oleh model menjadi lebih baik. Namun, berdasarkan percobaan yang dilakukan, data yang telah melalui proses pembersihan (cleaning) terlebih dahulu memberikan hasil yang terbaik. Oleh karena itu, sebelum memperbarui model, sebaiknya data diproses atau dibersihkan terlebih dahulu untuk memperoleh hasil yang maksimal.

Model yang dihasilkan tentu dapat berubah jika dilatih ulang, karena menggunakan prinsip neural network. Namun, yang pasti adalah lebih baik menggunakan model yang telah dilatih sebelumnya dan kemudian diperbarui, daripada membuat model dari awal, karena model yang diperbarui akan lebih kaya akan informasi yang dibutuhkan. Proses pelatihan model dan percobaan seperti analogi tentunya melibatkan parameter-parameter tertentu, sehingga percobaan dengan berbagai parameter, baik saat pelatihan maupun fine-tuning, akan mempengaruhi hasil akhir. Eksperimen dengan berbagai parameter sangat diperlukan untuk menemukan model yang terbaik.