Topic Modeling

Rabu, 3 Januari 2024

- Satria Bagus Panuntun

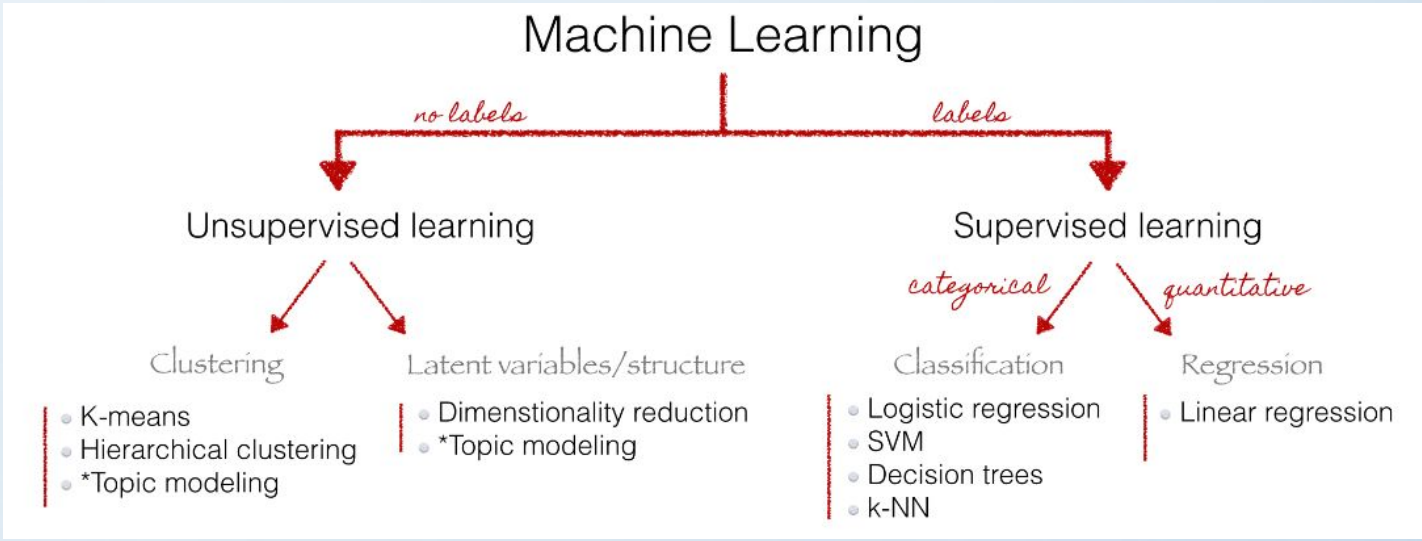
Isi

- Konsep topic modeling
- Tujuan topic modeling
- LDA
- Tahapan-tahapan topic modeling
- Praktik preprocessing

Konsep Dasar

Apa itu topic modeling?

Topic modeling merupakan salah satu pendekatan pada Text Mining yang cukup handal dalam melakukan penemuan data-data teks yang tersembunyi dan menemukan hubungan antara teks yang satu dengan lainnya dari suatu corpus (Jelodar, et al., 2018).



Apa itu topic modeling?

- Topic modeling adalah jenis model statistik untuk menemukan "topik" abstrak yang terdapat dalam kumpulan dokumen [1]
- Topic modeling adalah serangkaian algoritma yang mengungkap struktur tematik tersembunyi dalam koleksi dokumen. Algoritma ini membantu mengembangkan cara baru untuk mencari, menelusuri, dan meringkas arsip teks berukuran besar [2]
- Topik modeling menyediakan cara sederhana untuk menganalisis teks tak berlabel dalam jumlah besar. Sebuah "topik" terdiri dari sekelompok kata yang sering muncul bersamaan[3]

Apa itu topic modeling?

Mengelompokkan data teks berdasarkan suatu topik tertentu. Cara kerja topik modelling seperti clustering, dikatakan seperti clustering karena mengelompokkan dokumen berdasarkan kemiripannya. Topic Modelling termasuk unsupervised learning karena data yang digunakan tidak memiliki label.

Salah satu metode Topic Modeling adalah dengan menggunakan metode **Latent Dirichlet Allocation (LDA)**

Tujuan Topic Modeling

Mengidentifikasi Pola dalam Data Teks

- Menemukan topik tersembunyi dalam kumpulan dokumen besar.
- Contoh penelitian:
  1. Analisis tren topik dalam artikel berita selama satu dekade.
  2. Studi tentang topik yang sering dibahas dalam forum online.

Mengelompokkan Dokumen Berdasarkan Topik

- Mengelompokkan dokumen yang memiliki kesamaan topik.
- Contoh penelitian:
  1. Pengelompokan ulasan produk untuk analisis sentimen.
  2. Klasifikasi artikel jurnal berdasarkan bidang penelitian.

Menyederhanakan Data Teks untuk Analisis Lebih Lanjut

- Mengurangi dimensi data teks untuk memudahkan analisis.
- Contoh penelitian:
  1. Penggunaan topic modeling untuk menyederhanakan data teks sebelum analisis machine learning.
  2. Penerapan dalam analisis data besar untuk efisiensi.

Membantu dalam Rekomendasi Konten

- Menyediakan rekomendasi konten yang relevan berdasarkan topik yang diminati pengguna.
- Contoh penelitian:
  1. Sistem rekomendasi artikel berdasarkan topik yang sering dibaca pengguna.
  2. Rekomendasi film atau buku berdasarkan preferensi topik.

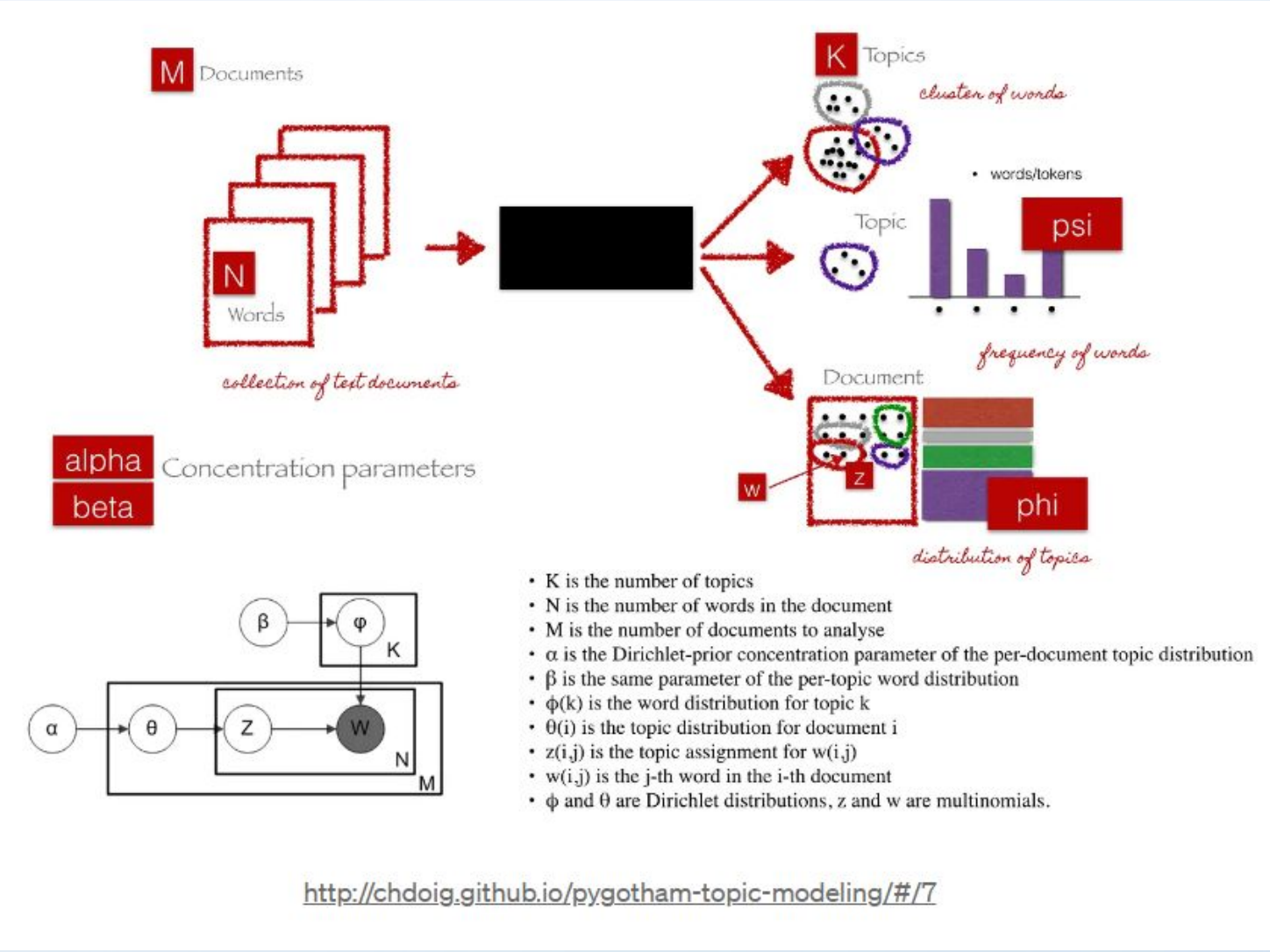
LDA

Latent Dirichlet Allocation (LDA)

LDA pertama kali diperkenalkan oleh Blei, Ng dan Jordan pada tahun 2003, adalah salah satu metode paling populer dalam pemodelan topik.

Putra & Kusumawardani (2017) mengatakan bahwa LDA dapat digunakan untuk meringkas, melakukan klasterisasi, menghubungkan maupun memproses data yang sangat besar karena LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen.

Latent Dirichlet Allocation (LDA)



http://chdoig.github.io/pygotham-topic-modeling/#/7

Komponen Utama:

- Dokumen (M): Kumpulan teks yang dianalisis.
- Kata (N): Kata-kata dalam dokumen.
- Topik (K): Kumpulan kata yang sering muncul bersama dalam dokumen.

Parameter Penting:

- α (alpha): Parameter konsentrasi untuk distribusi topik per dokumen.
- β (beta): Parameter konsentrasi untuk distribusi kata per topik.
- ψ (psi): Distribusi topik untuk dokumen tertentu.
- φ (phi): Distribusi kata untuk topik tertentu.

Kelebihan dan Kekurangan LDA

Kelebihan:

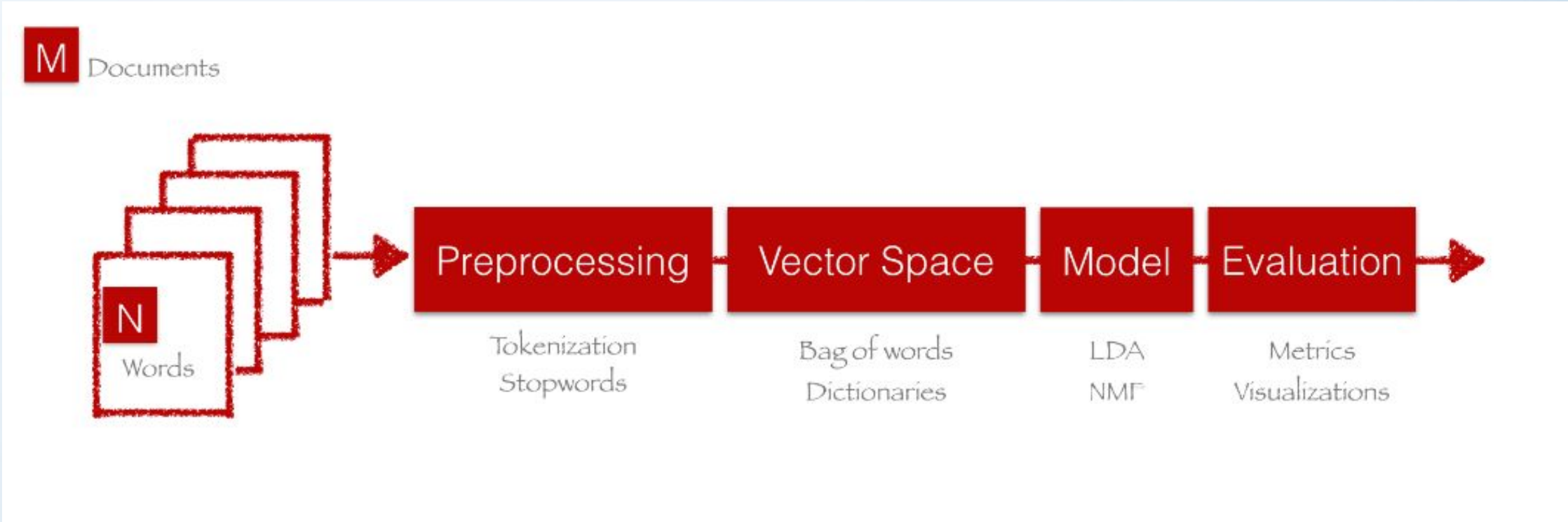
- Mampu menangani data teks besar.
- Menghasilkan interpretasi yang mudah dipahami dari data yang kompleks.
- Fleksibel dalam berbagai aplikasi, seperti analisis sentimen dan pengelompokan dokumen.

Kekurangan:

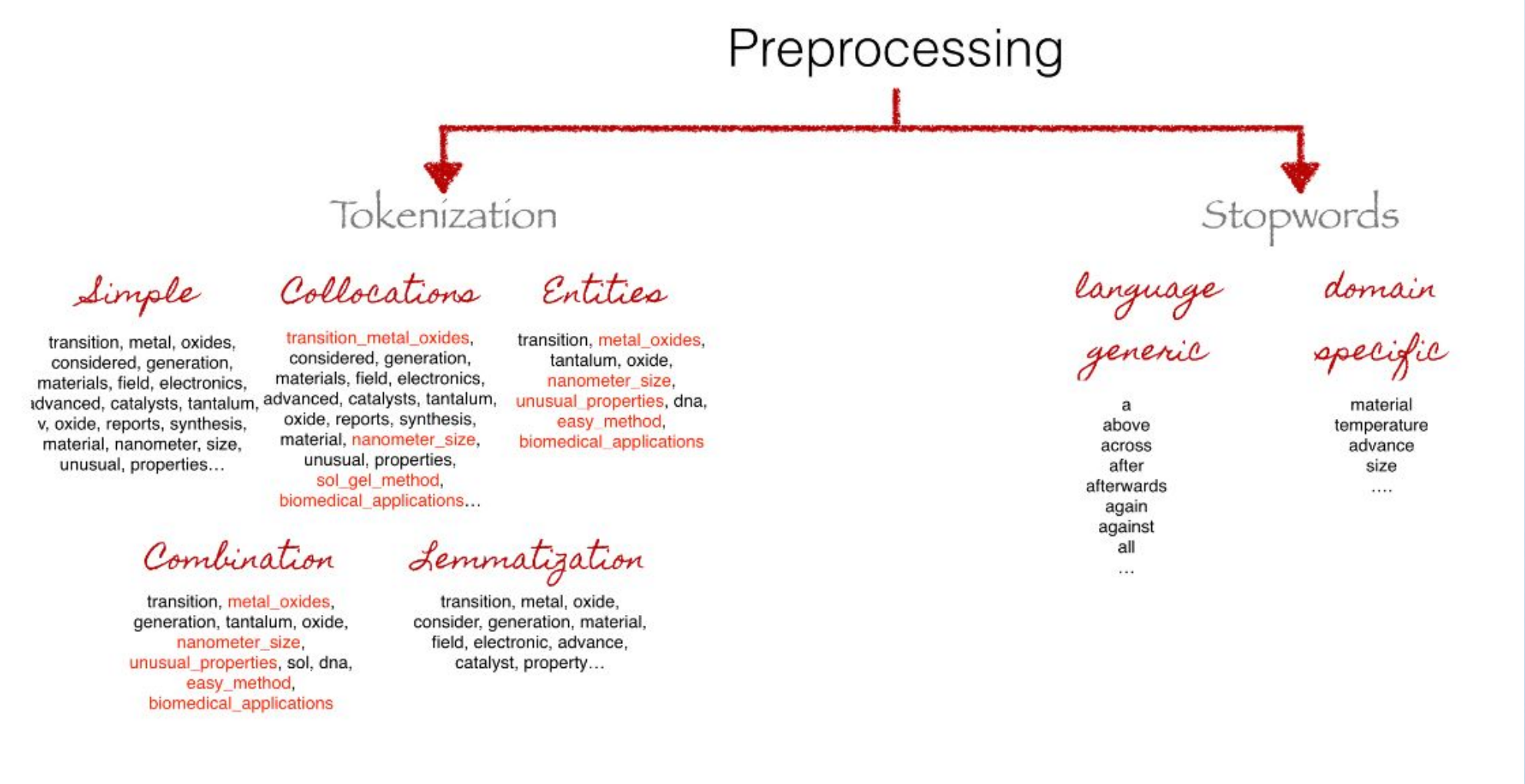
- Memerlukan pemilihan jumlah topik yang tepat.
- Sensitif terhadap parameter awal dan dapat terjebak dalam solusi lokal.
- Tidak mempertimbangkan urutan kata dalam dokumen.

Tahapan Topic Modeling

Tahapan topic modeling



Preprocessing

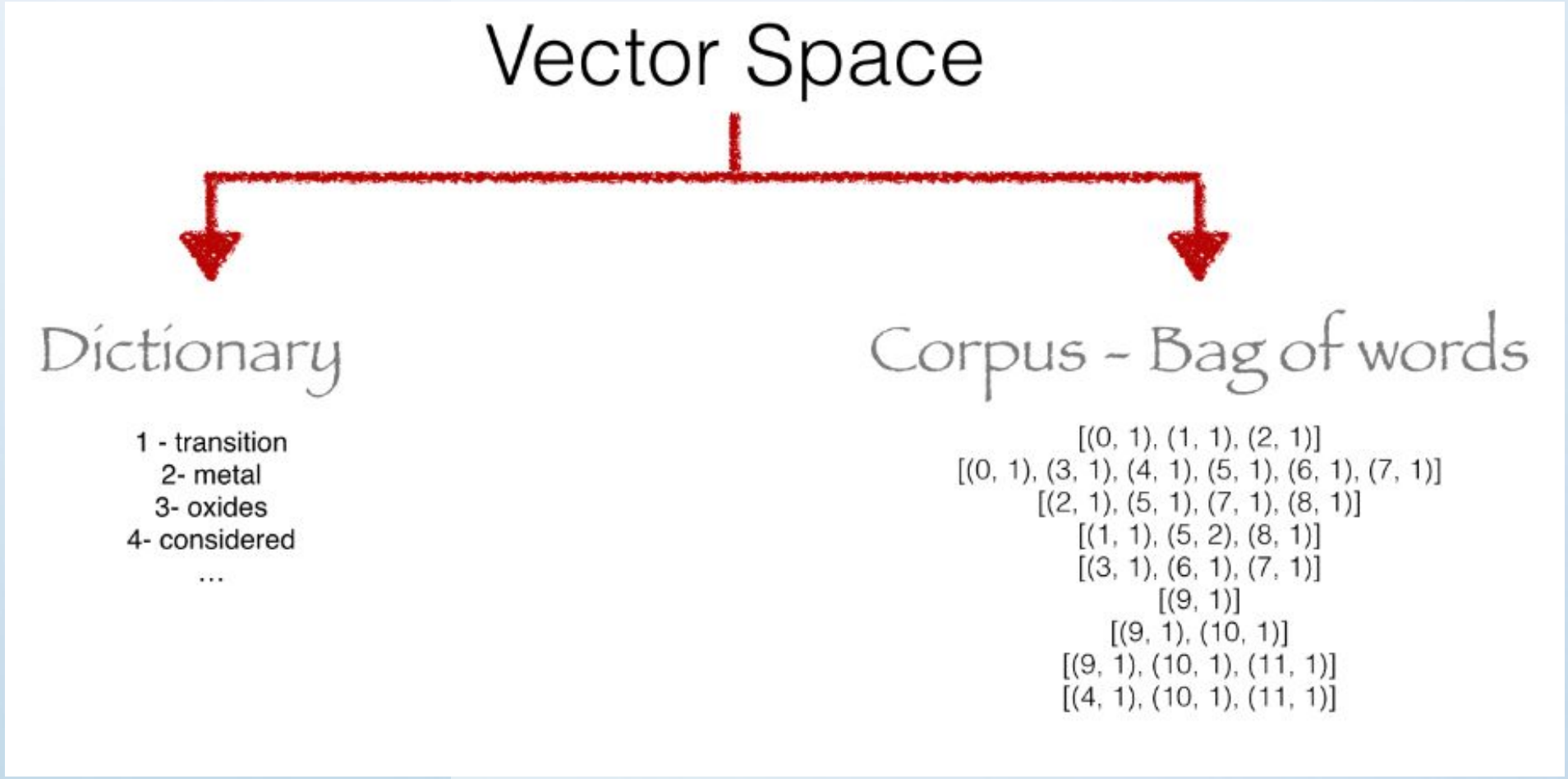


Preprocessing

Pertanyaan
Jajanan Pasar yang murah meriah dimana sih?
oleh-oleh jogja apa aja?
makan cupcake yang enak dimana ya?
Yg jual lemper di jogja mana ya?
Makanan ringan khas yogyakarta dimana??
Dimana yang menjual jajanan hits yang enak dan murah?
jual kue ulang tahun dimana?
Bakpia enak di jogja
Lokasi penjual jajanan oleh asli khas jogjakarta
Dimana lokasi menjual kue coklat lumer
kue artis di jogja apa aja?
makanan yang unik dan enak dimna sih? jual kue pancong dimna ? jual kembang tahu dimana saja ?
Toko Bakeries dimana ya?
makanan untuk oleh-oleh apa ya?
Tempat nongkrong yang asik dimana?
Tempat nongkrong yang harganya terjangkau dimana sih ?
Signature dish di kopi klotok apa ya ?
Tempat makan yg enak buat tugas dan ada wifi deket jakal dimana ya?

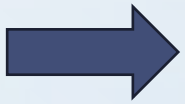
1	text
2	jajanan pasar murah meriah
3	
4	makan cupcake enak
5	lemper
6	ringan khas
7	jajanan hits enak murah
8	kue ulang tahun
9	bakpia enak
10	lokasi penjual jajanan asli khas jogjakarta
11	lokasi kue coklat lumer
12	kue artis

Vector Space



Vector Space

tweet_clean
selamat malam sobat industri kabar nih ekonomi nasional
alhamdulillah
emang top deh era presiden ri catat jaran nkri tumbuh ekonomi trwulan ii pandemi ekonomi tumbuh lejit
terimakasih obrol mas saran kritik makna upaya tingkat potensi sektor tani ekonomi indonesia



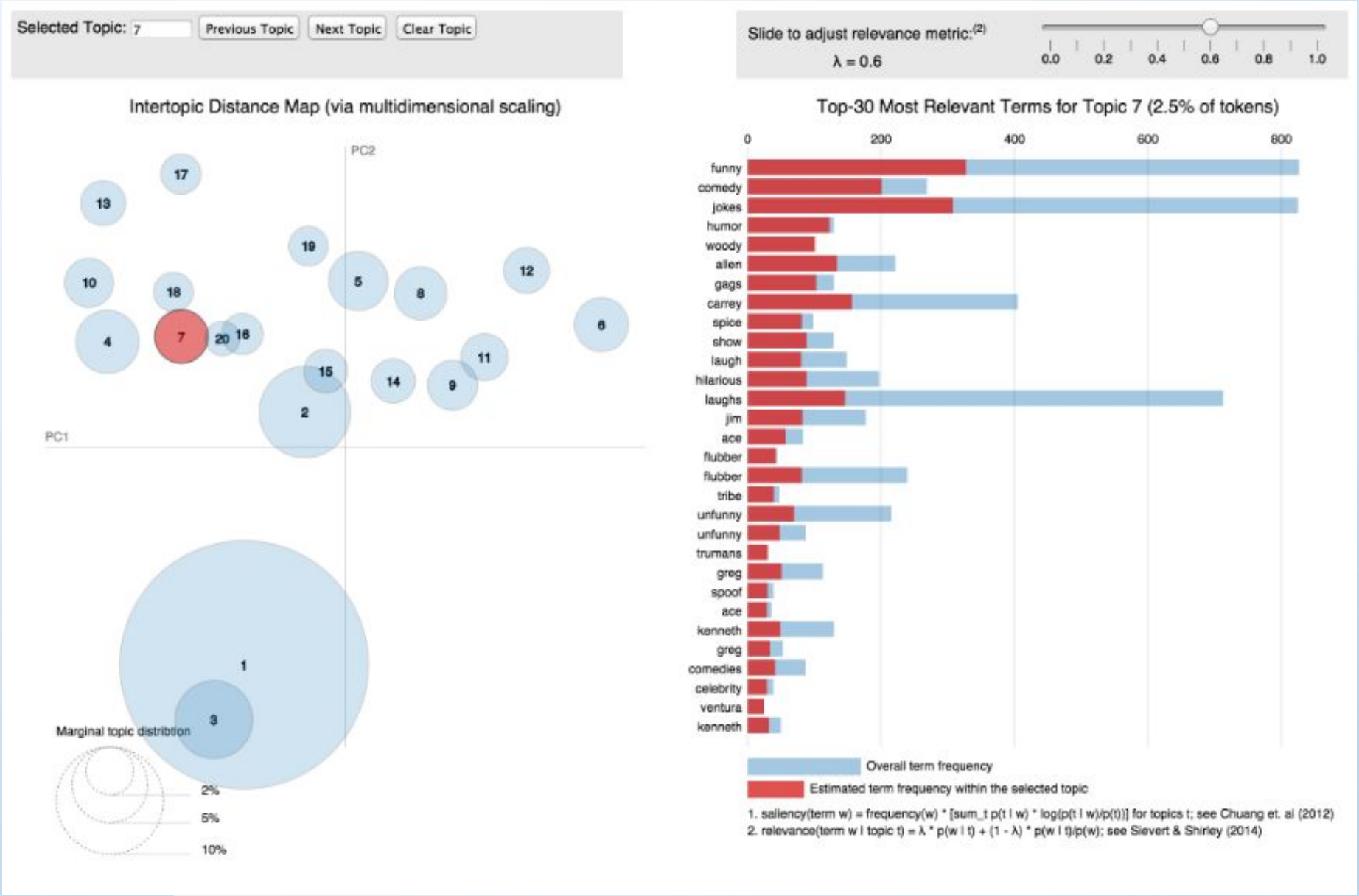
```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)]
[(5, 1)]
[(6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1)]
[(16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1)]
```

Gensim Model

Gensim Models

- `models.ldamodel` – Latent Dirichlet Allocation
- `models.ldamulticore` – parallelized Latent Dirichlet Allocation
- `models.lsimodel` – Latent Semantic Indexing
- `models.tfidfmodel` – TF-IDF model
- `models.rpmodel` – Random Projections
- `models.hdpmodel` – Hierarchical Dirichlet Process
- `models.logentropy_model` – LogEntropy model
- `models.lsi_dispatcher` – Dispatcher for distributed LSI
- `models.lsi_worker` – Worker for distributed LSI
- `models.lda_dispatcher` – Dispatcher for distributed LDA
- `models.lda_worker` – Worker for distributed LDA
- `models.word2vec` – Deep learning with word2vec
- `models.doc2vec` – Deep learning with paragraph2vec
- `models.phrases` – Phrase (collocation) detection
- `models.wrappers.ldamallet` – Latent Dirichlet Allocation via Mallet
- `models.wrappers.dtmmodel` – Dynamic Topic Models (DTM) and Dynamic Influence Models (DIM)
- `models.wrappers.ldavowpalwabbit` – Latent Dirichlet Allocation via Vowpal Wabbit

Evaluasi


Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.6
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
17
13
19
10
18
5
8
12
6
4
7
20
16
15
14
9
11
2
1
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (2.5% of tokens)
0
200
400
600
800
funny
comedy
jokes
humor
woody
allen
gags
carrey
spice
show
laugh
hilarious
laughs
jim
ace
flubber
flubber
tribe
unfunny
unfunny
trumans
greg
spoof
ace
kenneth
greg
comedies
celebrity
ventura
kenneth
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Evaluasi (Interpretasi)

```
Topic: 0 Word: 0.103*"makan" + 0.075*"halal" + 0.063*"pedas" + 0.053
*"korea" + 0.052*"lokasi" + 0.045*"murah" + 0.036*"lokasinya" + 0.034
*"menu" + 0.031*"cafe" + 0.027*"nyaman"
Topic: 1 Word: 0.153*"murah" + 0.116*"lokasi" + 0.069*"harga" + 0.064
*"seafood" + 0.048*"makan" + 0.039*"mahasiswa" + 0.029*"kantong" + 0.0
28*"terkenal" + 0.025*"terjangkau" + 0.025*"terenak"
Topic: 2 Word: 0.149*"khas" + 0.054*"masakan" + 0.052*"lokasi" + 0.046
*"promo" + 0.043*"wisata" + 0.041*"makan" + 0.034*"ayam" + 0.034*"nong
krong" + 0.032*"terdekat" + 0.029*"warung"
```

Ini berarti 10 kata kunci teratas yang berkontribusi pada topik 0 adalah: ‘makan’, ‘pedas’, ‘halal’, ‘korea’ .. dan seterusnya dan bobot ‘halal’ pada topik 0 adalah 0,075. Bobot mencerminkan betapa pentingnya kata kunci untuk topik itu.

Evaluasi (Interpretasi)

```
Topic: 0 Word: 0.103*"makan" + 0.075*"halal" + 0.063*"pedas" + 0.053
*"korea" + 0.052*"lokasi" + 0.045*"murah" + 0.036*"lokasinya" + 0.034
*"menu" + 0.031*"cafe" + 0.027*"nyaman"
Topic: 1 Word: 0.153*"murah" + 0.116*"lokasi" + 0.069*"harga" + 0.064
*"seafood" + 0.048*"makan" + 0.039*"mahasiswa" + 0.029*"kantong" + 0.0
28*"terkenal" + 0.025*"terjangkau" + 0.025*"terenak"
Topic: 2 Word: 0.149*"khas" + 0.054*"masakan" + 0.052*"lokasi" + 0.046
*"promo" + 0.043*"wisata" + 0.041*"makan" + 0.034*"ayam" + 0.034*"nong
krong" + 0.032*"terdekat" + 0.029*"warung"
```

Topik 0 memiliki model dengan pembahasan masakan Korea, halal, memiliki rasa pedas. Tidak hanya itu,topik ini juga membahas cafe yang nyaman, bagus, serta kekinian. Maka kelompok topik 2 isinya yaitu tentang kuliner/masakan luar negeri yang memiliki sertifikat halal di Yogyakarta.

Praktik Topic Modeling

Install Library NLTK

Natural Language Toolkit atau disingkat NLTK, adalah libray python untuk bekerja de
mempersiapkan teks sebelum digunakan pada machine learning atau algoritma dee

```
[1]  pip install nltk

     Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-pa
     Requirement already satisfied: click in /usr/local/lib/python3.10/dist-p
     Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-
     Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3
     Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-pa
```

Install Library Sastrawi

Python Sastrawi adalah pengembangan dari proyek PHP Sastrawi. Python Sastrawi
berimbuhan bahasa Indonesia menjadi bentuk dasarnya.

Link colab:

https://drive.google.com/drive/folders/13aQgM07VBb8Fafnz4GDC1AFzCvi9ETjd?hl=id