# Pembobotan *Vector Space Model* Korpus Twitter Tentang *Data Science Text Mining Text Retrieval* Menggunakan *Cosine Smiliarity*

**Muhammad Reza 2019470055**

**Selamet Saputra 2019470069**

**Syechan Ahmad Zidan 2019470110**

## Latar belakang masalah

*Text Mining* adanya *preprocessing* dan ekstraksi fitur, tahap *preprocessing* terdiri dari *case folding,stop word removal, stemming,word normalization* untuk mengatasi *overfitting* dari hasil *stemming* (Ma'rifah et al., 2020). ekstrasi fitur dalam text mining ada *term frequency* melihat setiap kata yang muncul didalam dokumen atau korpus, untuk menghitung *inverse document* diperlukan masing-masing kemunculan *term frequency* di setiap document atau korpus lalu dikalkulasi dengan rumus *Inverse Document Frequency* (IDF) .Maka dari hasil nilai ekstrasi fitur tf-idf ini digunakan untunk perhitung similaritas, dan untuk beberapa metode dalam pendekatan statistika (Setyawan et al., 2021).

## Identifikasi masalah

1. Melakukan *preprocessing document* teks
2. Melakukan ekstraksi fitur teks menggunakan *Term Frequency Inverse Document Frequency*
3. Menghitung jarak teks dokumen menggunakan *Cosine Smiliarity*

**Rumusan masalah**

1. Bagaimana melakukan preprocessing data teks *document* ?
2. Bagaimana cara melakukan ekstraksi fitur text dari *document* menggunakan *Term Frequency Inverse Document Frequency* ?
3. Bagaimana cara menghitung jarak teks *document* twitter menggunakan *Cosine Smiliarity* ?

**Batasan masalah**

1. Mengolah teks *document* dari twiiter tentang *data science text mining text retrieval*
2. Menggunakan bantuan *library* regex,NLTK, Pandas, Numpy untuk mengolah text data dan menghitung hasil jarak kedekatan dokument
3. Menggunakan data dari *twitter* berjumlah 84 dokument hanya bahasa inggris
4. Mengetahui hasil kedekatan jarak antara *document* menggunakan *cosine smiliarity*

## Korpus

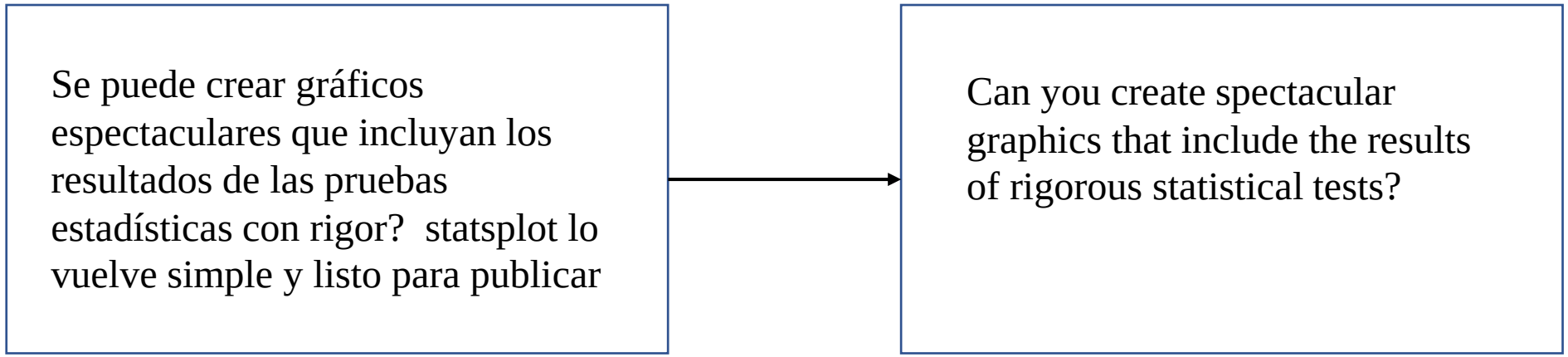
```
0     I think the message in Data Science needs to b...
1     Python libraries for:\n\n— Machine Learning\n—...
2                     Free Data Science PDF Books  \n🧵:
3     Top tech skills for a #DataEngineer in 2022 👨...
4     💡¿Se puede crear gráficos espectaculares que i...
                            ...
59    Excellent retrieval skills in #BusheyHeathRead...
60    A novel adapter-based method for parameter-eff...
61    On @jhuclsp YouTube: Changes in Tweet Geolocat...
62    Yes, I am looking for a summer 2023 research i...
63    trec: TREC collection (2010). A bipartite netw...
Name: full_text, Length: 64, dtype: object
```

```
0     Every story in the world has one of 6 basic pl...
1     SoLA invites you to a lecture on “Text Mining ...
2     Check out our events happening this week! \n\n...
3     The RuMOR team is growing! Thanks to @SSHRC_CR...
4     I'm doing a lot of preaching right now to coll...
5     Why my #Geosis package is simply the most robu...
6     Text Mining and Analytics #TextMining https://...
7     Let's speed up my booming Twitter career! Here...
8     Are you after a course that will teach you the...
9     Fundamentals of Predictive Text Mining (Texts ...
10    Awesome strategies for our humanities courses ...
11                          meaning of life is number 42
12    Brisbane Data, Power BI and AI Bootcamp speake...
13                               this is a possible tweet
14                                        It's a Tweet!
15                               this is an example tweet
16                                this is your next tweet
17                              or, maybe, a possible badger
18    Python Text Mining: Perform Text Processing, W...
19           and now for something completely different
Name: full_text, dtype: object
```

Data yang digunakan  adalah data atau korpus yang diambil berasal dari twitter dengan cara *scrapping*, korpus yang diambil adalah tentang text mining dan *information retrieval*. Berjumlah 84 dokument

## Data Exploration

Karena Batasan masalah diperuntukkan hanya untuk bahasa inggris, pada saat dilakukan pencarian korpus untuk data, terdapat kalimat non bahasa inggris, maka dilakukan transformasi menjadi bahasa inggris



## Data Exploration II

Terdapat 347 stopwords di korpus dengan menggunakan *library nltk*

| *Stopwords* | *Frequensi* |
|---|---|
| to | 53 |
| in | 46 |
| and | 46 |
| a | 44 |
| the | 38 |
| of | 33 |
| is | 26 |
| for | 24 |
| you | 19 |
| on | 18 |

## Preprocessing

Dalam tahap ini, dilakukan berupa :

- Case folding
- Stopword removal
- Stemming
- Word normalization

I think the message in Data Science needs to be: Don't believe everything you read. 🧵

#stats #datascience https://t.co/4jGMgmX8Nw



think message data science needs believe everything read stats datascience jgmgmx nw

# Term Frequency

Setelah dilakukan preprocessing maka dilakukan *Term frequency*, yaitu pemecahan kalimat menjadi kata di setiap dokumen serta mencari kemunculan kata dari masing-masing dokumen

| Term(kata) | Dokument frequency | Frequency in dokument |
|---|---|---|
| data | 23 | 1 |
| text | 44 | 1 |
| mining | 27 | 1 |
| retrieval | 18 | 1 |
| python | 11 | 1 |

## *Term Frequency Inverse Document Frequency Query*

Pada tahap ini dilakukan pencarian kueri untuk melihat hasil yang muncul dari kueri yang diinput

| Term | DF | N/df | IDF |
|---|---|---|---|
| data | 23 | 3.652173913043478 | 0, 5625514500442887 |
| text | 44 | 1.9090909090909092 | 0.2808266095756942 |
| mining | 27 | 3.1111111111111111 | 0.49291552190289434 |
| retrieval | 18 | 4.6666666666666667 | 0.6690067809585756 |
| python | 11 | 7.6363636363636637 | 0.8828866009036567 |

# *Vector space model*

- Sesudah dilakukan tf-idf maka dilakukan *vector space model* menggunakan *cosine smiliarity* untuk melihat jarak kueri dengan dokumen

| Query | Dokument terkuat | Hasil smiliaritas |
|---|---|---|
| data text mining retrieval python | 20 | 0.1838423837595189 |