## **Petunjuk Pengerjaan Case Study**

Pada tugas ini, Anda akan diberikan dua buah korpus yang dapat digunakan, yaitu berkas **Korpus.json** dan **Korpus\_Ringkas.json**. Berkas **Korpus.json** yang diberikan merupakan berkas yang berisi informasi lengkap dari berita dalam bahasa Indonesia yang terdiri dari:

| **Kolom** | **Deskripsi** |
| --- | --- |
| section | Kategori berita seperti ‘ekonomi’, ‘sosial’, ‘internasional’, dst. [String] |
| articleid | ID dari berita [int 6 bit / digit] |
| articletype | Tipe penampilan artikel seperti ‘singlepage’ [String] |
| createdate | Tanggal dipublikasikannya sebuah berita [YYYY-MM-DD HH:MM:SS] |
| author | Penulis berita [String] |
| originalTitle | Judul berita [String] |
| tag | Penggolongan berita berdasarkan kata kunci tertentu [String] |
| keywords | Kata kunci utama berita [String] |
| subsection | Kategori lanjutan dari ‘section’ seperti ‘Peristiwa’ [String] |
| content | Konten utama dari berita yang disampaikan [String] |

Sedangkan berkas dari **Korpus\_Ringkas.json** merupakan potongan dari korpus lengkap yang telah disesuaikan, isi dari berkas tersebut terdiri dari:

| **Kolom** | **Deskripsi** |
| --- | --- |
| id | ID dari artikel [int] |
| content | Konten utama dari berita yang disampaikan [String] |

Pada tugas ini Anda diminta untuk melakukan analisis korpus menggunakan korpus yang telah diberikan. Untuk soal di mana Anda diminta untuk menampilkan histogram dan *word cloud*, Anda **diperbolehkan** menggunakan *library* tambahan. (cth: matplotlib, wordcloud, etc.). Untuk proses segmentasi kalimat dan tokenisasi kata, Anda **diwajibkan** menggunakan library nltk.

Berikan dokumentasi pada implementasi Anda. ***Deliverable*** akhir merupakan **1 berkas jupyter notebook (dalam format ipynb)**, di mana pada berkas tersebut terdapat kode yang Anda tuliskan + dokumentasi (bisa berupa chart, teks, atau lainnya) + penjelasan terkait apa yang telah Anda kerjakan.

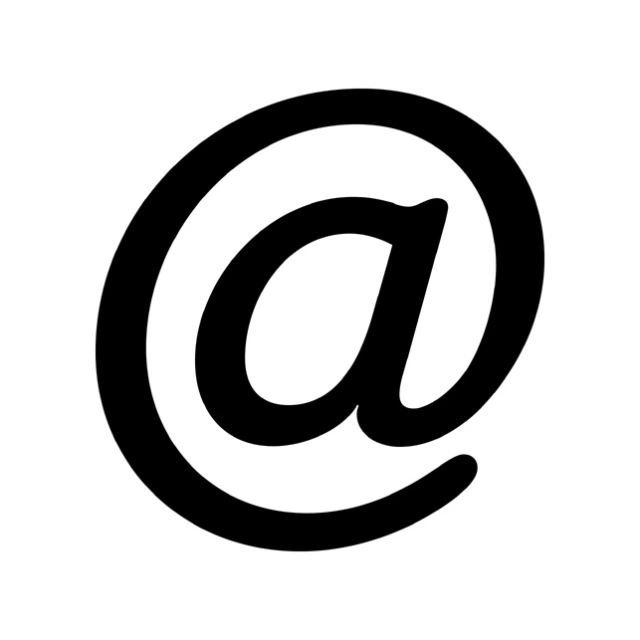
Untuk bagian A dan B, gunakan data yang terdapat pada *file* “**Korpus.json**”. Kemudian, untuk bagian C dan D, gunakan data pada *file* “**Korpus\_Ringkas.json**”.

## 

## **A - Corpus Statistics (25 Poin)**

1. **[1]** Berapakah jumlah data yang terdapat pada “**Korpus.json**”?
2. **[6]** Pada soal ini, lakukan analisa untuk fitur “section” pada korpus dengan menjawab pertanyaan-pertanyaan berikut ini!
   1. **[2]** Hitunglah berapa banyak jumlah “section” yang **unik**. Sebutkan “section” apa saja yang terdapat dalam korpus tersebut!
   2. **[4]** Tampilkan jumlah data untuk setiap “section” dalam bentuk **histogram**!
3. **[12]** Pada soal ini, lakukan analisa untuk fitur “content” pada korpus dengan menjawab pertanyaan-pertanyaan berikut ini! Untuk mempermudah pengerjaan, suatu token dalam korpus dapat dianggap sebagai kata jika seluruh karakter berupa ***alphanumeric*** (A-z, 0-9).
   1. **[3]** Hitunglah jumlah kata yang **unik** dalam korpus. Catatan: Kata “Jakarta” dan “jakarta” dianggap sebagai kata yang sama.
   2. **[5]** Tampilkan 100 kata diurutkan berdasarkan jumlah kemunculan terbanyak dalam bentuk **histogram**! Kemudian, sebutkan 10 kata terbanyak dalam korpus beserta jumlahnya!
   3. **[4]** Sebutkan 10 kata **bigram** terbanyak dalam korpus!
4. **[6]** Pada soal ini, lakukan analisa untuk fitur “originalTitle” pada korpus dengan menjawab pertanyaan-pertanyaan berikut ini! Identik dengan soal sebelumnya, suatu token dalam korpus dapat dianggap sebagai kata jika seluruh karakter berupa ***alphanumeric*** (A-z, 0-9).
   1. **[2]** Berapa **rata-rata** jumlah kata yang terdapat dalam “originalTitle”?
   2. **[4]** Tampilkan ***word cloud*** untuk fitur “originalTitle” dalam korpus!

## **B - Regex (25 Poin)**

****

Salah satu simbol yang unik dan sering digunakan secara langsung dalam teks adalah “*at symbol*”. Pada awalnya, karakter “@” umumnya digunakan pada bidang akuntansi atau penulisan *invoice* sebagai pengganti istilah “*at a rate of*” atau dengan kata lain menyatakan harga satuan suatu barang. Saat ini, karakter “@” lebih sering digunakan pada akun Email dan akun media sosial seperti Twitter, Instagram, TikTok, dan lainnya.

Sebagai calon IR engineer yang mahir, tentunya diharapkan kita dapat membedakan penggunaan karakter “@” ini dalam konteks yang berbeda-beda. Menggunakan dataset “**Korpus.json**”, khususnya fitur “content” pada korpus tersebut, **tentukan** apakah setiap karakter “@” yang ada dalam korpus **termasuk dalam penggunaan untuk jenis** akun Email, Twitter, atau Instagram!

Anda diharuskan untuk merancang **regex** (boleh lebih dari satu) untuk menjawab soal ini. Gunakan *guideline* berikut ini untuk membantu Anda. Asumsikan bahwa seluruh penggunaan karakter “@” pada korpus ini **hanya** termasuk pada jenis akun Email, Twitter, atau Instagram (tidak ada media sosial atau penggunaan lainnya).

**Ketentuan Akun Email / Twitter / Instagram**

1. Setiap akun pastinya hanya memiliki satu karakter “@”
2. Setiap akun juga tidak dapat diakhiri dengan tanda “.”
3. Akun Email memiliki karakter “@” yang diapit oleh dua token. Dengan kata lain, suatu akun Email berbentuk [Token1]@[Token2].
   1. Setiap token pada suatu akun Email dapat terdiri dari karakter-karakter ***alphanumeric*** (A-z, 0-9), tanda “.”, tanda “-”, atau tanda “\_”
   2. Khusus untuk Token 2, harus terdapat minimal 1 tanda “.”
   3. Beberapa contoh akun email **valid**:
      1. akun-ir@gmail.com
      2. akun\_\_ir@cs.ui\_\_
      3. ir.4.life@my-hobbies.uk.id-
   4. Beberapa contoh akun email **tidak valid**:
      1. not$valid\_@cs.ui
      2. m4ta\_bat1n@magician\_\_
      3. raiden-shogun-op@genshin.life.
4. Akun media sosial memiliki karakter “@” di awal dan diikuti oleh satu token. Dengan kata lain, suatu akun media sosial berbentuk @[Token].
   1. Token maksimal terdiri dari 15 karakter
   2. Token pada akun media sosial hanya bisa terdiri dari karakter-karakter ***alphanumeric*** (A-z, 0-9), tanda “.”, atau tanda “\_”
   3. Beberapa contoh akun media sosial **valid**:
      1. @iwuvu3000
      2. @the\_king.ar
      3. @1panic\_\_
   4. Beberapa contoh akun media sosial **tidak valid**:
      1. @info-retrieve
      2. @lulus2021.
      3. @fake\_soc\_med\_account
   5. Khusus untuk akun Twitter, dapat dipastikan bahwa **tidak** terdapat karakter “.” pada token

**Soal**:

1. **[10]** Tuliskan **seluruh akun Email** yang terdapat pada korpus tersebut dan **jumlah** setiap akun Email tersebut **muncul** dalam korpus!
2. **[9]** Tuliskan **seluruh akun media sosial** yang terdapat pada korpus tersebut dan **jumlah** setiap akun media sosial tersebut **muncul** dalam korpus!
3. **[6]** Berdasarkan jawaban dari nomor 2, tentukan akun yang **dapat dipastikan** merupakan akun Instagram (harus dengan regex untuk menentukannya)!

**Catatan**:

Jawaban Anda untuk bagian ini diharuskan berbentuk baris-baris yang berisi “[NAMA\_AKUN] - [JUMLAH\_MUNCUL]” untuk setiap akun yang ditemukan. Contoh:

* Untuk akun Email:

perolehan\_informasi@cs.ui.ac.id - 2

genshin-anniversary@mihoyo.com - 6

* Untuk akun media sosial:

@naed\_aar - 4

@dota2 - 3

## **C - Tokenization (15 Poin)**

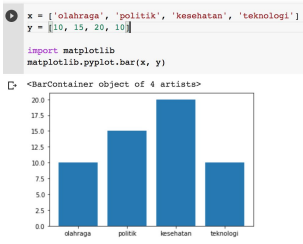
1. **[2]** Transformasi semua karakter dalam “**Korpus\_Ringkas.json**” ke format ***lowercase***!
2. **[3]** **Hilangkan** seluruh karakter berupa angka, whitespace berlebih, dan tanda baca!
3. **[6]** Lakukan **tokenisasi** pada setiap baris data. Tampilkan hasil tokenisasi berupa ***list of tokens*** dari 5 data pertama!
4. **[4]** Untuk setiap baris data tentukan token yang dapat dikategorikan sebagai ***stopwords***! Hilangkan kata-kata tersebut dari *list of tokens*!

## **D - Stemming dan Lemmatization (35 Poin)**

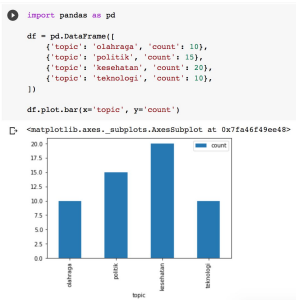
1. **[11]** Menggunakan *list of tokens* yang diperoleh dari bagian C, lakukan ***stemming*** dengan menggunakan ***library* Sastrawi** untuk masing-masing token pada setiap baris data. Kemudian, buatlah pemetaan antara token sebelum di-*stem* dengan token setelah di-*stem* menggunakan Sastrawi **(*before stem & after stem*)**. Cukup tampilkan pemetaan pada token dari 5 data pertama saja.
2. **[8]** Menggunakan *list of tokens* yang diperoleh dari bagian C, lakukan ***lemmatization*** untuk masing-masing token pada setiap baris data menggunakan ***library* Stanza**. Kemudian, tampilkan hasil *list of tokens* dari 5 data pertama!

## **Lampiran**

* Sastrawi: <https://github.com/har07/PySastrawi>
* Stanza: <https://stanfordnlp.github.io/stanza/>
* Spacy: <https://spacy.io/usage>
* Contoh menampilkan histogram dengan *library* matplotlib



* Contoh menampilkan histogram dari data frame Pandas



* Contoh menampilkan *word cloud* dengan library [wordcloud](https://pypi.org/project/wordcloud/)

