PROJECT AKHIR STKI - E

STOPWORD ANALISIS DENGAN METODE ZIPF LAW, MUTUAL INFORMATION, DAN NLTK STOPWORD

Presented by kelompok 6

WHAT WE DO ?

FETCH DATA

https://www.kaggle.com/nltkdata/web-text-corpus)

INSTALL LIBRARY

numpy, pandas,
sklearn, math, nltk

PREPROCESSING

case folding, cleansing,
splitting data,
removing stopword

RANKING EVALUATION

cosine similarity,
precision @k, MAP

PREPROCESSING

Case folding

https://www.kaggle.com/nltkdata/web-text- corpus)

Splitting data

15 datas class firefox & 15 datas class overheard

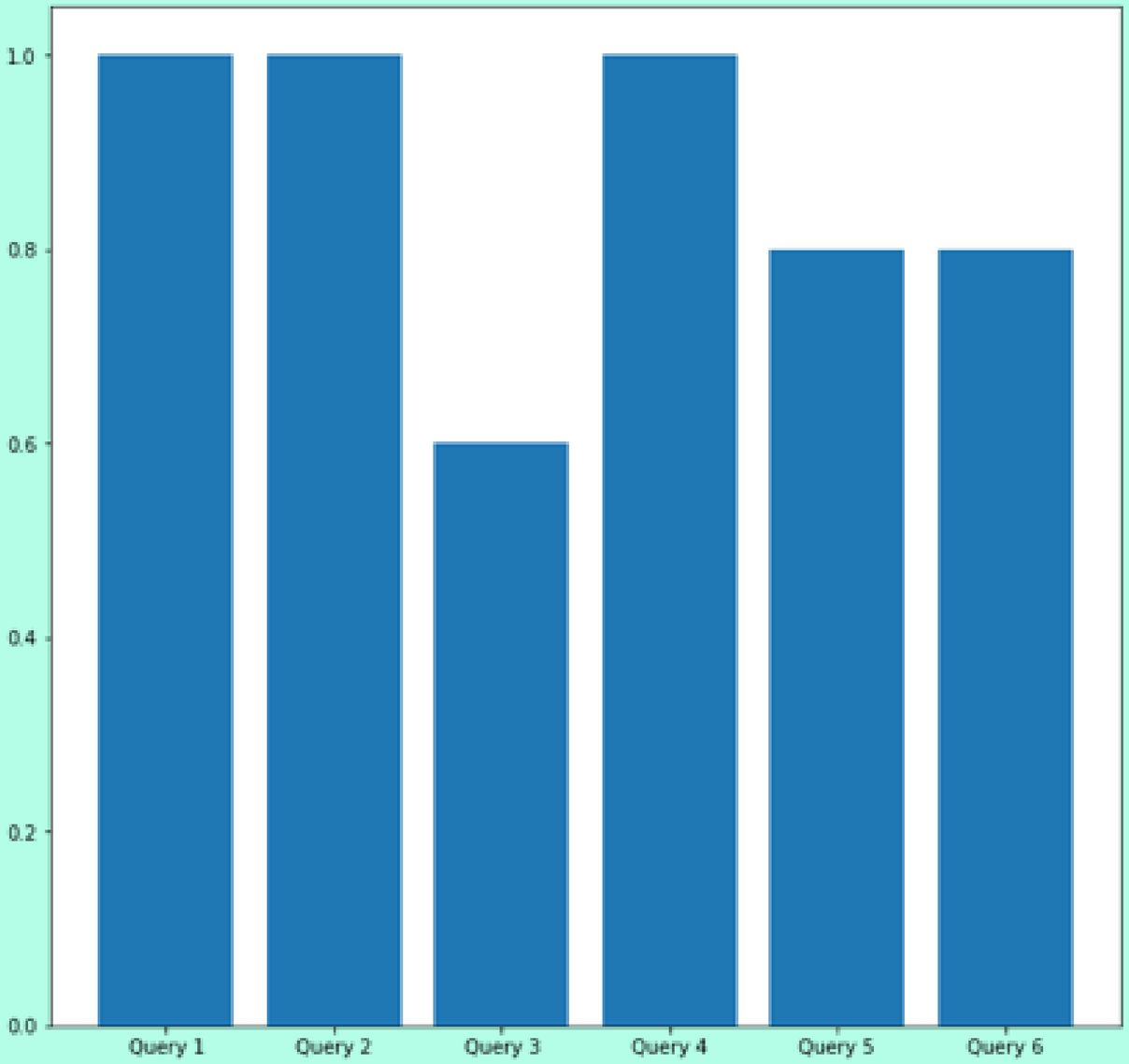
Cleansing

removing punctuations, html, url

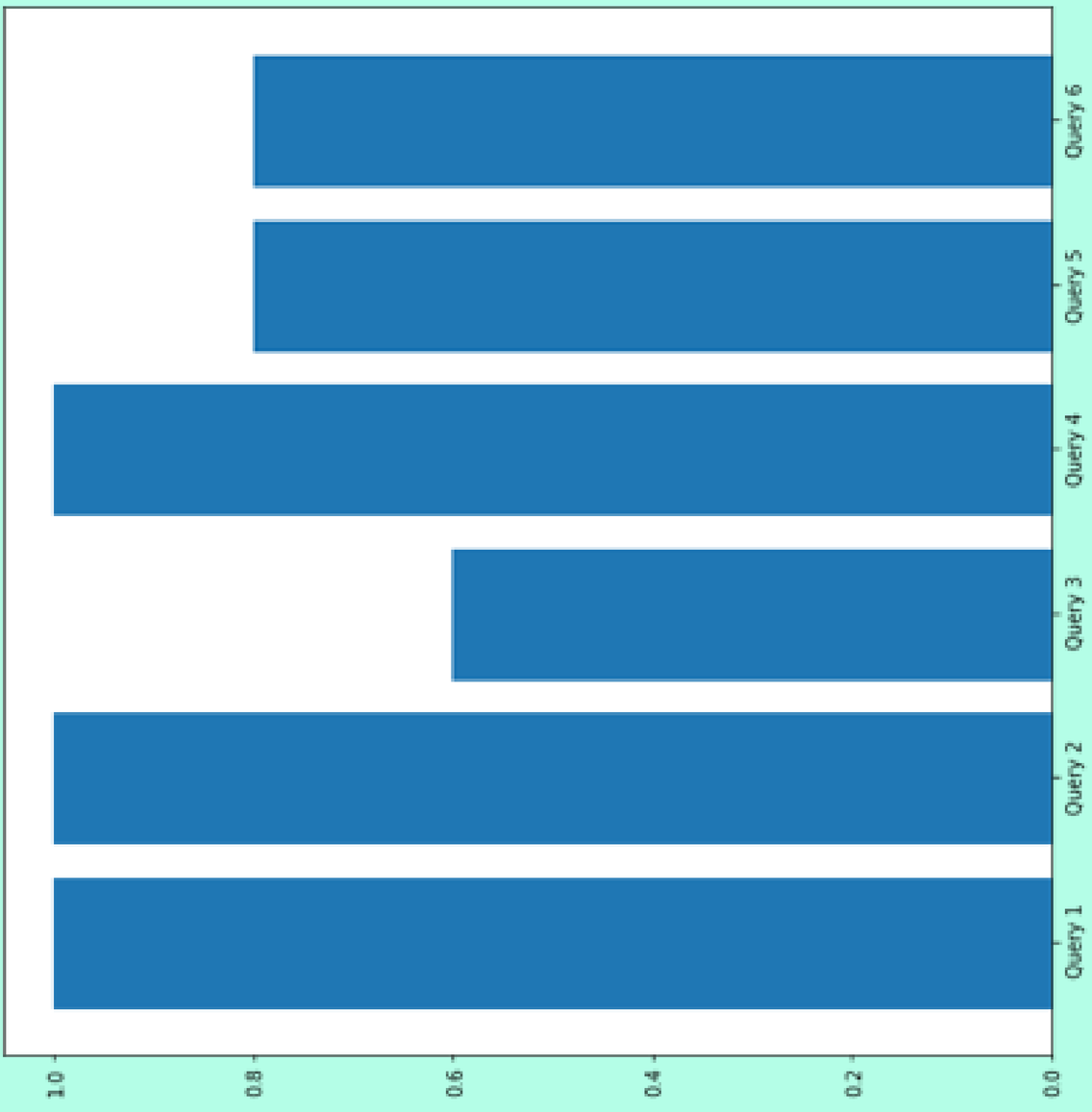
Remove stopword

Zipf law, mutual information, nltk stopword

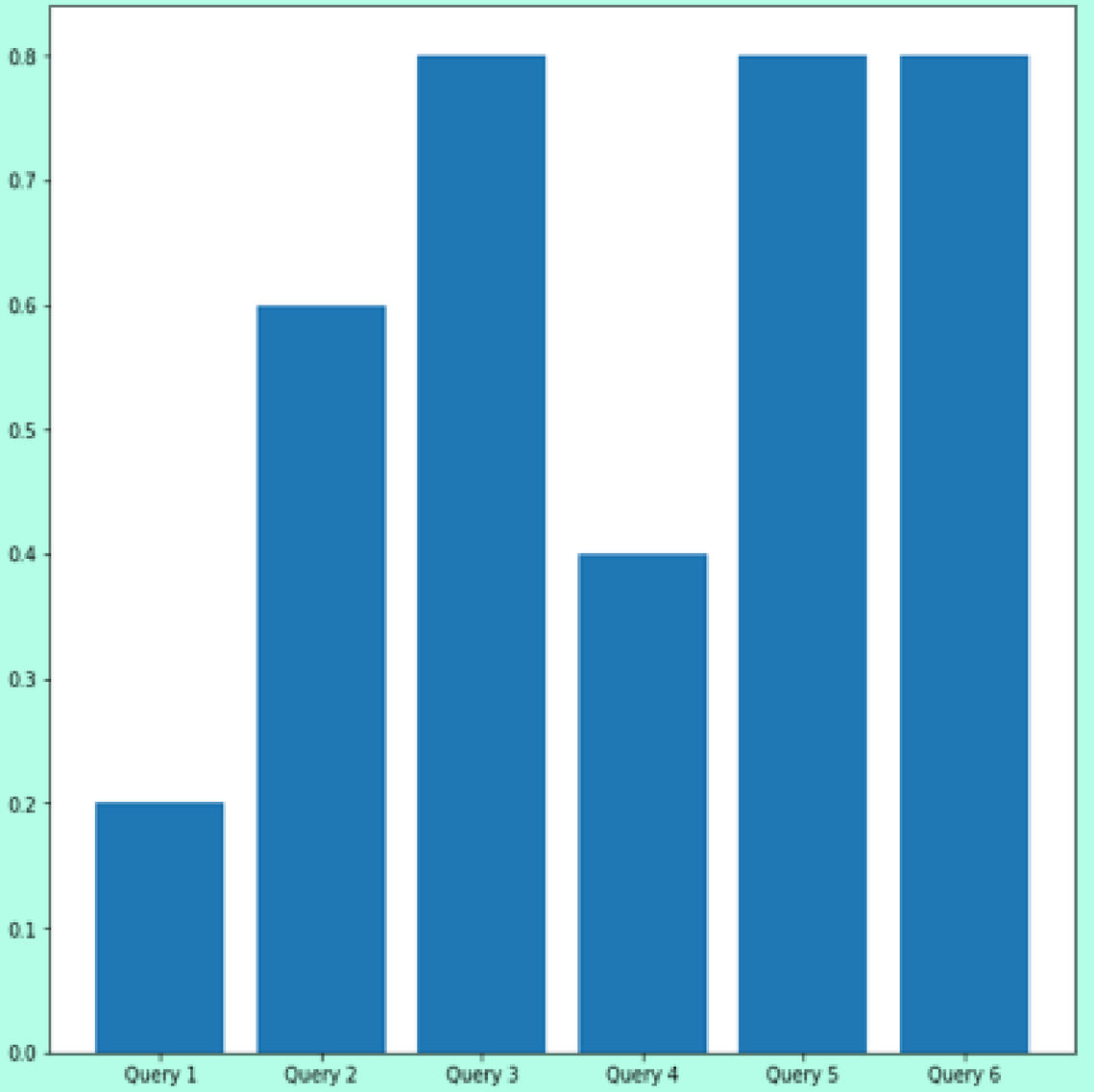
Precision @K NLTK



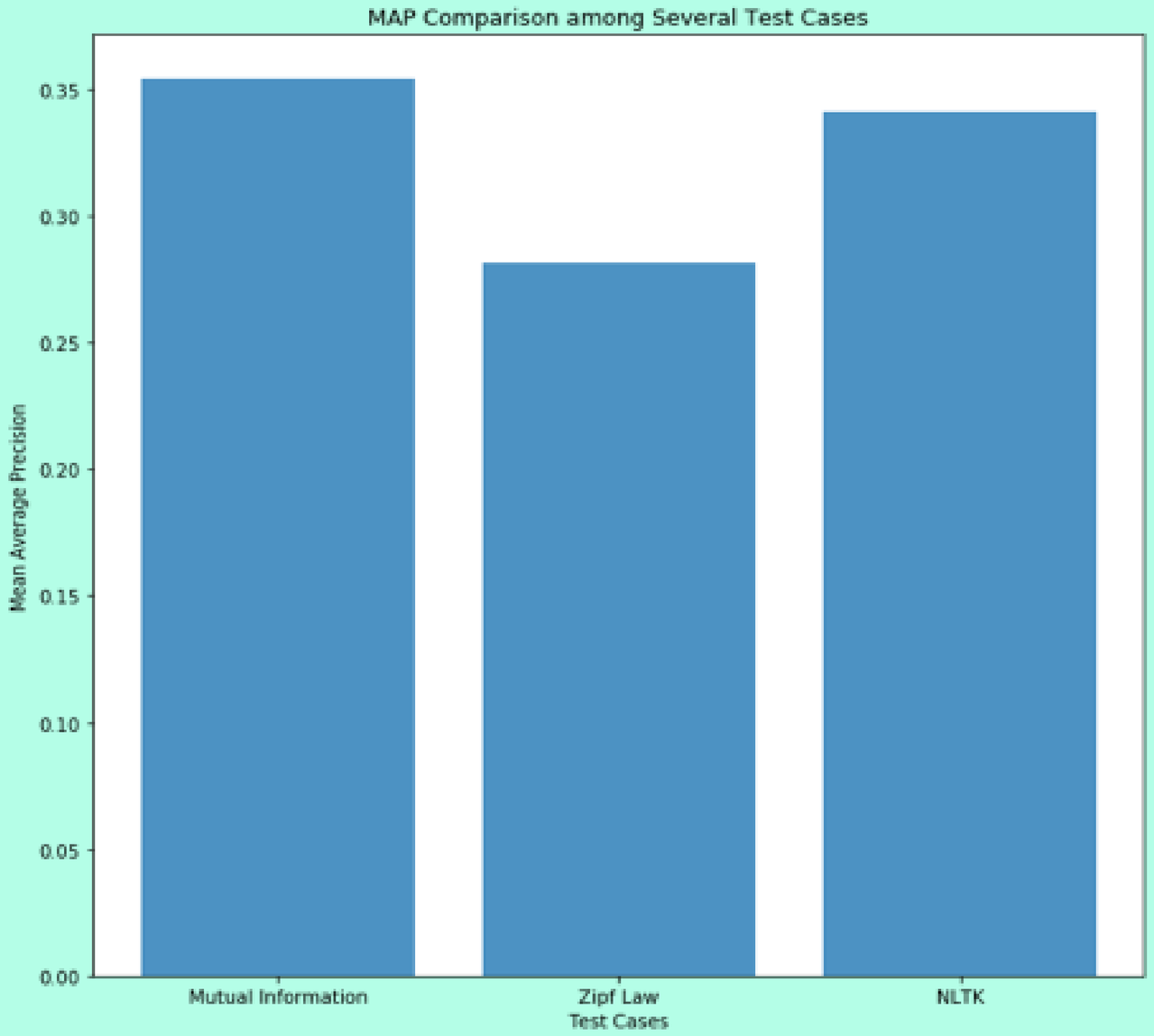
Precision @K Mutual Information



Precision @K Zipf Law



Mean Average Precision



Conclusion

DARI PERCOBAAN STUDI KASUS INI DIDAPATKAN BAHWA , MAP YANG DIHASILKAN DARI TIAP METODE BAIK MENGGUNAKAN NLTK , MI & ZIPF LAW RENDAH , HAL INI DIKARENAKAN CORPUS YANG DIGUNAKAN MEMILIKI JUMLAH KATA YANG SEDIKIT DI SETIAP DOKUMEN NYA . DAN JIKA DILIHAT DARI HASIL EVALUASI MENGGUNAKAN PRECISION@K , MAKA METODE YANG TERBAIK YANG DIGUNAKAN UNTUK MENENTUKAN STOPWORD ADALAH NLTK DAN MUTUAL INFORMATION , KARENA JIKA DILIHAT DARI GRAFIK , KECENDERUNGAN QUERY YANG MEMILIKI TINGKAT PRECISION@K YANG RENDAH KEBANYAKAN ADA JIKA MENGGUNAKAN METODE ZIPF LAW , HAL INI DIAKIBATKAN KARENA BEBERAPA QUERY MEMILIKI DOKUMEN RELEVANT YANG SEDIKIT , DARI JUMLAH YANG DITENTUKAN YAITU LIMA DOKUMEN TERATAS , BERDASARKAN NILAI COSSIM DENGAN QUERY YANG DIPILIH.

Suggestion

- Menambah jumlah data yang digunakan untuk corpus , dengan jumlah lebih dari 15

- Menggunakan proses Stemming

Thank You

questions are welcome.

FARHAN SETYA DHITAMA

Bagian Dokumentasi
(Judul , Mengatur Bagi Tugas , membuat PPT)

LUDGERUS DARELL

Bagian mencari Data ,Dokumentasi bagian Hasil Analisis, Kesimpulan dan Saran

MAHENDRA OKZA PRADHANA

Semua Coding Selain PreProcessing , Saran

TITUS CHRISTIAN

Coding preProcessing