

# Natural Language Processing - Bigram

Nama: Helmi Satria Nugraha

Nim: 1301154325

## Keterangan sumber artikel, topik, dan alasan pemilihan

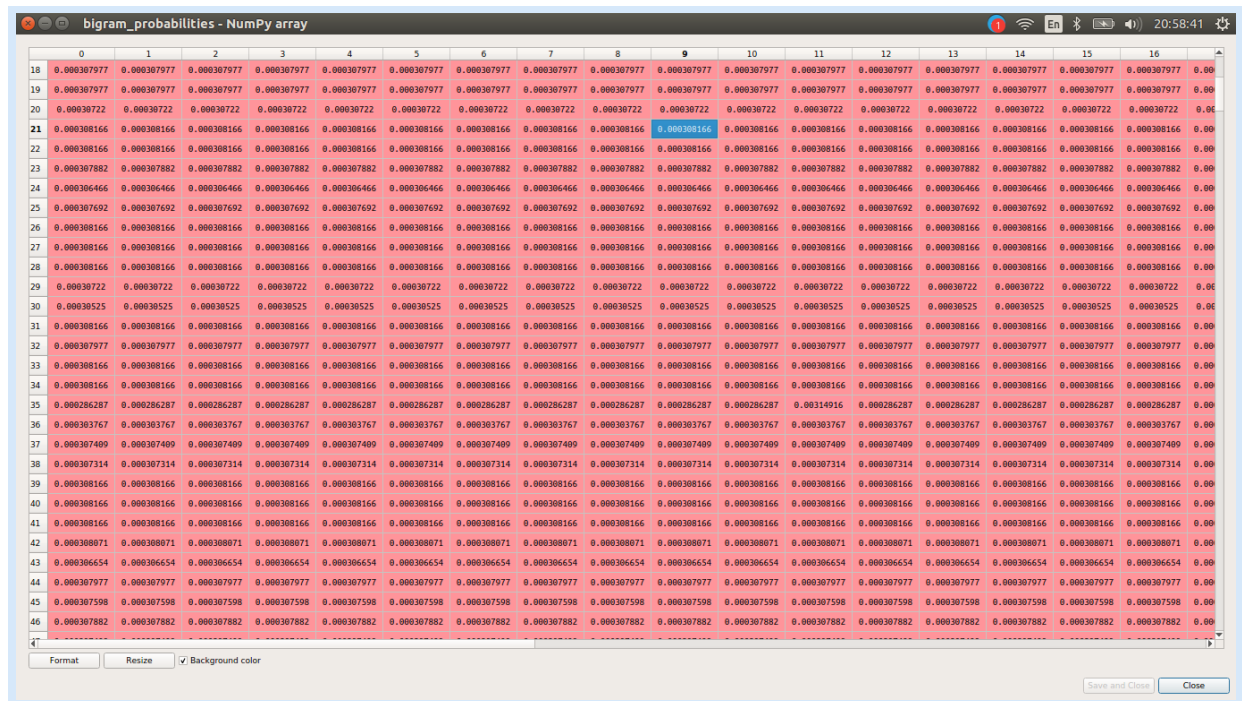
Sumber artikel: <http://katadata.co.id/indeks/search/-/-/-/255/->

Topik: Katadata - Keuangan

Alasan pemilihan: Katadata mempunyai indeks listing dan mudah untuk di *crawl* karena memiliki fitur pemilihan topik, fitur *pagination* dan katadata menyediakan indeks **listing keseluruhan** bukan indeks berita per hari.

## Analisis terhadap hasil pengujian prediksi kemunculan kata

Next word akan menunjukkan prediksi kata (yang sudah ada di dataset) berikutnya. Prediksi kemunculan kata berikutnya dari input suatu kata didapatkan dari tabel **bigram\_possibilities** dimana tabel tersebut didapatkan dari tabel **bigram\_counts**. Kata yang dipilih adalah pasangan kata yang memiliki probabilitas tertinggi.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
18	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977
19	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977
20	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722
21	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
22	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
23	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882
24	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466	0.000306466
25	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692	0.000307692
26	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
27	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
28	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
29	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722	0.00030722
30	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525	0.00030525
31	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
32	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977
33	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
34	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
35	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287	0.000286287
36	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767	0.000303767
37	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409	0.000307409
38	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314	0.000307314
39	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
40	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
41	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166	0.000308166
42	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071	0.000308071
43	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654	0.000306654
44	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977	0.000307977
45	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598	0.000307598
46	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882	0.000307882

Bigram Possibilities

```
In [32]: runfile('/home/helmisatria/Documents/Kuliah/NLP/bigram-natural-language-processing/bigram/input.py', wdir='/home/helmisatria/Documents/Kuliah/NLP/bigram-natural-language-processing/bigram')

Masukkan kata: bencana
next word: bnpb

Masukkan kata: utang
next word: korporasi

Masukkan kata: negara
next word: bumh

Masukkan kata: bank
next word: indonesia

Masukkan kata: uang
next word: muka

Masukkan kata: otoritas
next word: jasa

Masukkan kata: pajak
next word: apabila

Masukkan kata: |
```

*Prediksi kata berikutnya*

## Analisis terhadap hasil evaluasi perplexity

Berikut adalah 5 kalimat yang saya pilih:

1. Nasabah pemegang kartu debit Standard Chartered Bank Indonesia
2. Pemegang Nasabah debit kartu Standard Chartered Indonesia Bank
3. Otoritas Jasa Keuangan OJK meluncurkan Paket Kebijakan Agustus
4. Jasa Otoritas Keuangan meluncurkan OJK Paket Agustus Kebijakan
5. Kementerian Pekerjaan Umum dan Perumahan Rakyat PUPR akan memperketat pengawasan

Kalimat 1, 2 dan 3,4 adalah **kalimat yang sama** hanya saja dilakukan proses **reorder** sehingga dapat mengetahui perbedaan atau perbandingan nilai perplexity jika suatu kalimat yang sama memiliki **tata urutan kata yang berbeda**.

Mengatur ulang atau mengacak tata urutan kata dalam suatu kalimat dapat didapatkan nilai perplexity nya jika sudah dilakukan proses **add-one** atau **leplace smoothing** karena jika ada pasangan kata yang tidak memiliki bigram akan bernilai 0 dan jika pembagi sama dengan 0 akan terjadi error dalam perhitungan perplexity atau probability

Kalimat 5 hanya tambahan jika suatu kalimat memiliki kata yang banyak untuk mengetahui seberapa besar dampak jumlah kata untuk nilai perplexity

Berikut adalah proses perhitungan perplexity tiap kalimat

```

.... perplexity, perplexity
Kalimat: ['nasabah', 'pemegang', 'kartu', 'debit', 'standard', 'chartered', 'bank', 'indonesia']
Sum of Log -50.61771591528034
Perplexity: 80.29367768224796
Kalimat: ['pemegang', 'nasabah', 'debit', 'kartu', 'standard', 'chartered', 'indonesia', 'bank']
Sum of Log -67.8024811255782
Perplexity: 355.89556404708173
Kalimat: ['otoritas', 'jasa', 'keuangan', 'ojk', 'meluncurkan', 'paket', 'kebijakan', 'agustus']
Sum of Log -52.99317496617109
Perplexity: 98.64314353269899
Kalimat: ['jasa', 'otoritas', 'keuangan', 'meluncurkan', 'ojk', 'paket', 'agustus', 'kebijakan']
Sum of Log -79.49429722120288
Perplexity: 980.1013637206648
Kalimat: ['kementerian', 'pekerjaan', 'umum', 'dan', 'perumahan', 'rakyat', 'pupr', 'akan', 'memperketat', 'pengawasan']
Sum of Log -80.29578852335123
Perplexity: 261.3028142300783

```

### *Perhitungan perplexity dari beberapa kalimat*

Dari proses perhitungan tersebut, didapatkan hasil sebagai berikut:

1. Semakin besar nilai **sum of log** semakin baik model sentence yang diperiksa
2. Semakin kecil nilai **perplexity** semakin baik model sentence yang diperiksa
3. Semakin panjang/banyak jumlah kata di suatu sentence semakin besar nilai perplexity