Nama : Muhamad Feriyanto

Kelas : 3SD1

NIM : 222011347

Responsi Information Retrieval

Pertemuan 9

Language Model for Information Retrieval

Query = "*vaksin corona jakarta*"

Corpus = berita

Standard Query Likelihood

Top 3 Standard Query Likelihood

In [78]:
```
rank_lm = exact_top_k_lm(no_sw, likelihood_scores,3)
for doc_id in rank_lm:
    print (f"{doc_id}: {rank_lm[doc_id]}")
```

```
berita3: 1.5996992565397702e-05
berita1: 0.0
berita2: 0.0
```

Rank Dokumen	Likelihood Score
Berita 3	1,76E-10
Berita 1	0,00
Berita 2	0,00

Dalam Standard Query Likelihood didapatkan rank 3 teratas pada dokumen 'Berita3', 'Berita1', dan 'Berita2'. Berita 3 memiliki likelihood score yang sangat kecil (mendekati 0), sedangkan berita lainnya likelihood scorenya adalah 0 ada metode Standard Query Likelihood.

Smoothing

Laplace Smoothing

Top 3 Laplace Smoothing

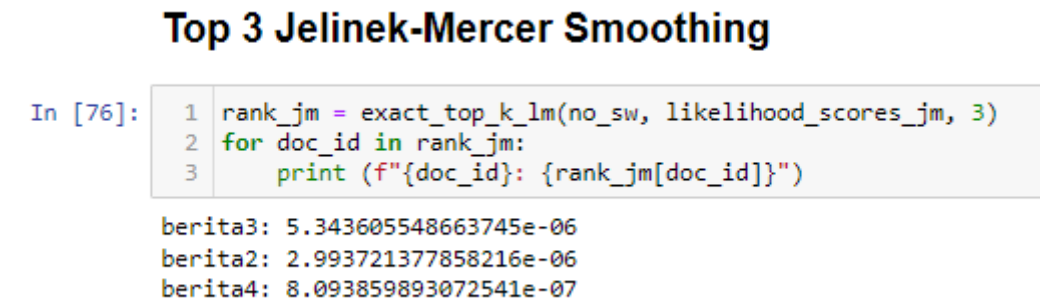
In [77]:
```
rank_lap = exact_top_k_lm(no_sw, likelihood_scores_lap, 3)
for doc_id in rank_lap:
    print (f"{doc_id}: {rank_lap[doc_id]}")
```

```
berita3: 2.870317203059586e-06
berita2: 1.7176733341735618e-06
berita4: 1.017708121310808e-06
```

Rank Dokumen	Likelihood Laplace Score
Berita 3	2.870317203059586E-06
Berita 2	1.7176733341735618E-06
Berita 4	1.017708121310808E-06

Dengan menggunakan Laplace Smoothing, didapatkan rank 3 teratas pada dokumen ‘Berita3’, ‘Berita2’, dan ‘Berita4’ dengan nilai yang mendekati 0.

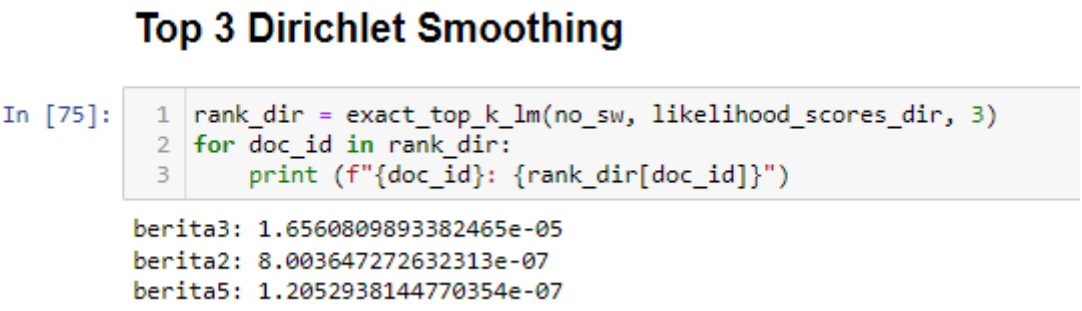
Jelinek-Mercer Smoothing



Rank Dokumen	Likelihood Jelinek-Mercer Score
Berita 3	5.343605548663745e-06
Berita 2	2.993721377858216e-06
Berita 4	8.093859893072541e-07

Dengan menggunakan Jelinek-Mercer Smoothing, didapatkan rank 3 teratas pada dokumen ‘Berita3’, ‘Berita2’, dan ‘Berita4’ dengan nilai yang mendekati 0.

Dirichlet Smoothing



Rank Dokumen	Likelihood Jelinek-Mercer Score
Berita 3	1.6560809893382465e-05
Berita 2	8.003647272632313e-07
Berita 5	1.2052938144770354e-07

Dengan menggunakan Dirichlet Smoothing, didapatkan rank 3 teratas pada dokumen ‘Berita3’, ‘Berita2’, dan ‘Berita5’ dengan nilai yang mendekati 0.

Tambahan:

Okapi BM25

In [77]:
```
from collections import OrderedDict
def exact_top_k_bm25(doc_dict, rank_score, k):
    relevance_scores = {}
    i = 0
    for doc_id in doc_dict.keys():
        relevance_scores[doc_id] = rank_score[i]
        i = i + 1

#         sorted_value = OrderedDict(sorted(relevance_scores.items(), key=lambda x: x[1], reverse = True))
#         top_k = {j: sorted_value[j] for j in list(sorted_value)[:k]}

        sorted_value = OrderedDict(sorted(relevance_scores.items(),key = lambda x:x[1], reverse = True))
        top_k = {j : sorted_value[j] for j in list(sorted_value)[:k]}
    return top_k
```

In [80]:
```
exact_top_k_bm25(no_sw, querinisasi("vaksin corona jakarta"), 3)
```

Out[80]:
```
{'berita3': 1.068210132282897,
 'berita2': 0.8332116429494796,
 'berita5': 0.544069042444546}
```

VSM

In [114]:
```
# Menyimpan skor kemiripan dalam suatu list
from collections import OrderedDict
def exact_top_k_vsm(doc_dict, TD, q,k):
    relevance_scores = {}
    i = 0
    for doc_id in doc_dict.keys():
        relevance_scores[doc_id] = cosine_sim(q, TD[:,i])
        i  = i + 1

    sorted_value = OrderedDict(sorted(relevance_scores.items(),key = lambda x:x[1], reverse = True))
    top_k = {j : sorted_value[j] for j in list(sorted_value)[:k]}
    return top_k
top_3_vsm = exact_top_k_vsm(no_sw, TD, TQ[:, 0], 3)
print(top_3_vsm)
```

```
{'berita2': 0.9937562849886573, 'berita3': 0.9932276247792821, 'berita1': 0.9918235081246563}
```

Tabel perbandingan

VSM, Okapi BM25, Standard Query Likelihood, Laplace Smoothing, Jelinek-Mercer Smoothing, Dirichlet Smoothing

Rank	Okapi BM25		VSM		Likelihood		Laplace		JM		Dirichlet	
	Doc	Nilai	Doc	Nilai	Doc	Nilai	Doc	Nilai	Doc	Nilai	Doc	Nilai
1	Berita3	1,0682	Berita2	0,9938	Berita3	1,76e-10	Berita3	2,87e-06	Berita3	5,34e-06	Berita3	1,65e-05
2	Berita2	0,8332	Berita3	0,9932	Berita1	0,00	Berita2	1,71e-06	Berita2	2,99e-06	Berita2	8,00e-07
3	Berita5	0,5441	Berita1	0,9918	Berita2	0,00	Berita4	1,01e-06	Berita4	8,09e-07	Berita5	1,21e-07

Kesimpulan:

Dalam beberapa metode yang telah diterapkan, terdapat perbedaan dalam dokumen yang di-*retrieve.*