# 最先进的语义搜索句子相似度计算

原创 ronghuaiyang AI公园 6月7日

作者：Daulet Nurmanbetov
编译：ronghuaiyang

导读

语义搜索是NLP中很值得去解决的，但又很困难的问题。

我们通常会花很多时间在大量的文档中寻找特定的信息。我们通常会使用CTRL + F。还有众所周知的Google-fu，在21世纪的职场中，有效使用Google搜索信息的是一项宝贵的技能。人类的所有知识对我们来说都是可用的，问题在于提出正确的问题，以及知道如何浏览结果找到相关的答案。

我们的大脑会执行语义搜索，我们会查看结果并找到与我们的搜索查询相似的句子。在金融和法律行业尤其如此，因为文件越来越长，我们不得不搜索很多关键字来找到正确的句子或段落。时至今日，人类在探索上所付出的累积努力是惊人的。

自NLP出现以来，机器学习一直试图解决语义搜索的这个问题。一个完整的研究领域 —— 语义搜索已经出现。最近，由于深度学习技术的进步，计算机能够以最小的人力投入精确地向我们提供相关信息。

## 句子嵌入的方法

自然语言处理(NLP)领域对此有一个术语，当一个词被提及时，我们称之为“surface form”，举个例子，“president”这个词本身意味着国家元首。但根据上下文和时间，这可能意味着特朗普或奥巴马。

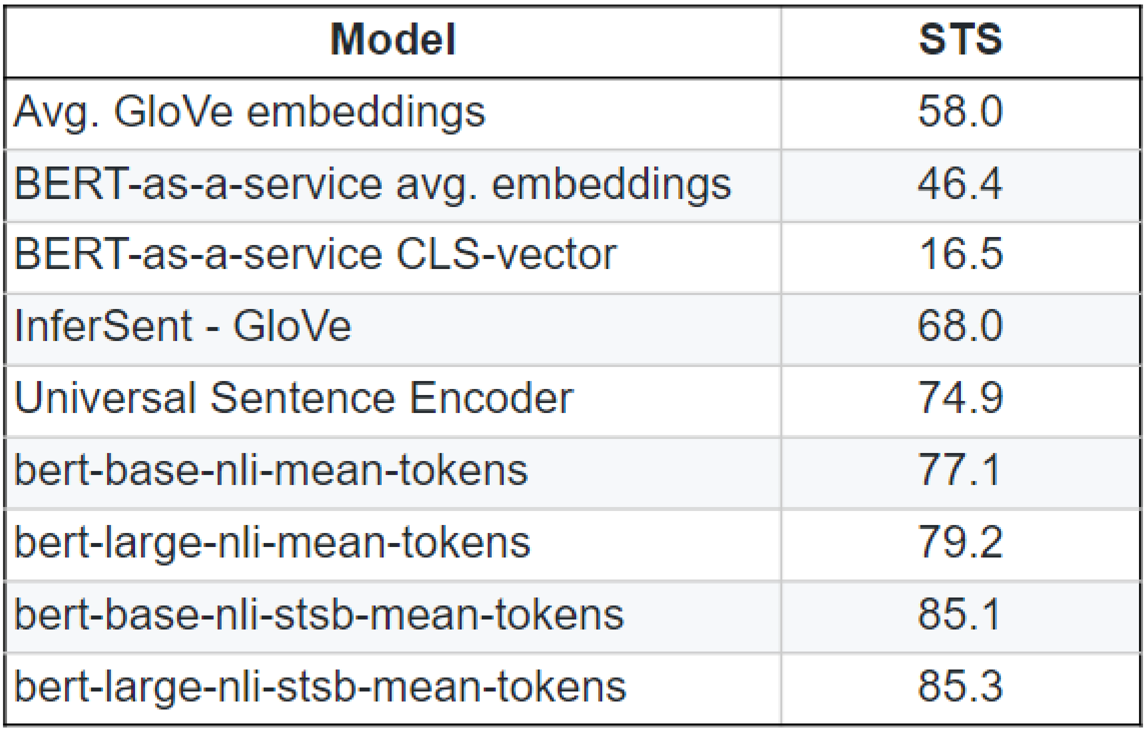
NLP的进步使我们能够有效地映射这些surface form，并将这些单词中的上下文捕获到称为“embeddings”的东西中。具有相似含义的两个单词将具有相似的向量，从而允许我们计算向量的相似性。

扩展这个想法，在向量空间中，我们应该能够计算任意两个句子之间的相似性。这就是句子嵌入模型所能达到的效果。这些模型将任何给定的句子转换成一个向量，从而能够快速计算任意一对句子的相似度或不同度。

## 最先进的语义搜索 —— 找到最相似的句子

这个想法并不新鲜，最早的一篇论文——word2vec早在2013年就提出了用向量表示单个单词。然而，从那时起，BERT和其他基于Transformer的模型让我们走了很长的路，它们允许我们更有效地捕捉这些词的上下文。

在这里，我们如何将最近的嵌入模型与word2vec或过去的GloVe进行比较。



| Model | STS |
|---|---|
| Avg. GloVe embeddings | 58.0 |
| BERT-as-a-service avg. embeddings | 46.4 |
| BERT-as-a-service CLS-vector | 16.5 |
| InferSent - GloVe | 68.0 |
| Universal Sentence Encoder | 74.9 |
| bert-base-nli-mean-tokens | 77.1 |
| bert-large-nli-mean-tokens | 79.2 |
| bert-base-nli-stsb-mean-tokens | 85.1 |
| bert-large-nli-stsb-mean-tokens | 85.3 |

STS是NLP的句子意义相似度竞赛。得分越高更好

这些经过修改和微调的BERT NLP模型在识别相似的句子方面非常好，比以前的模型好得多。让我们看看这在实际意义上意味着什么。

我有几篇2020年4月的文章标题，我希望找到与一组搜索词最相似的句子。

这里是我的搜索词 ——

```
1. The economy is more resilient and improving.
2. The economy is in a lot of trouble.
3. Trump is hurting his own reelection chances.
```

我的文章标题如下 ——

```
Coronavirus:
White House organizing program to slash development time for coronavirus vacci
Trump says he is pushing FDA to approve emergency-use authorization for Gilead
AstraZeneca to make an experimental coronavirus vaccine developed by Oxford Un
Trump contradicts US intel, says Covid-19 started in Wuhan lab. (The Hill)
Reopening:
Inconsistent patchwork of state, local and business decision-making on reopeni
White House risks backlash with coronavirus optimism if cases flare up again (
Florida plans to start reopening on Monday with restaurants and retail in most
California Governor Newsom plans to order closure of all state beaches and par
Japan preparing to extend coronavirus state of emergency, which is scheduled t
Policy/Stimulus:
Economists from a broad range of ideological backgrounds encouraging Congress 
Global economy:
China's official PMIs mixed with beat from services and miss from manufacturin
China's Beige Book shows employment situation in Chinese factories worsened in
Japan's March factory output fell at the fastest pace in five months, while re
Eurozone economy contracts by 3.8% in Q1, the fastest decline on record (FT)
US-China:
Trump says China wants to him to lose his bid for re-election and notes he is 
Senior White House official confident China will meet obligations under trad d
Oil:
Trump administration may announce plans as soon as today to offer loans to oil
Munchin says Trump administration could allow oil companies to store another s
Norway, Europe's biggest oil producer, joins international efforts to cut supp
IEA says coronavirus could drive 6% decline in global energy demand in 2020 (F
Corporate:
Microsoft reports strong results as shift to more activities online drives gro
Facebook revenue beats expectations and while ad revenue fell sharply in March
```

```
Tesla posts third straight quarterly profit while Musk rants on call about nee
eBay helped by online shopping surge though classifieds business hurt by closu
Royal Dutch Shell cuts dividend for first time since World War II and also sus
Chesapeake Energy preparing bankruptcy filing and has held discussions with le
Amazon accused by Trump administration of tolerating counterfeit sales, but co
```

在计算了每个查询和每个嵌入的相似性后，这里是我的每个搜索词的前5个相似的句子：

```
======================
Query: The economy is more resilient and improving.Top 5 most similar sentence
Microsoft reports strong results as shift to more activities online drives gro
Facebook revenue beats expectations and while ad revenue fell sharply in March
Senior White House official confident China will meet obligations under trad d
Economists from a broad range of ideological backgrounds encouraging Congress 
White House risks backlash with coronavirus optimism if cases flare up again (
======================
Query: The economy is in a lot of trouble.Top 5 most similar sentences in corp
Inconsistent patchwork of state, local and business decision-making on reopeni
eBay helped by online shopping surge though classifieds business hurt by closu
China's Beige Book shows employment situation in Chinese factories worsened in
Eurozone economy contracts by 3.8% in Q1, the fastest decline on record (FT) (
China's official PMIs mixed with beat from services and miss from manufacturin
======================
Query: Trump is hurting his own reelection chances.Top 5 most similar sentence
Trump contradicts US intel, says Covid-19 started in Wuhan lab. (The Hill) (Sc
Amazon accused by Trump administration of tolerating counterfeit sales, but co
Trump says China wants to him to lose his bid for re-election and notes he is 
Inconsistent patchwork of state, local and business decision-making on reopeni
White House risks backlash with coronavirus optimism if cases flare up again (
```

你可以看到，这个模型挑选出最相似的句子是多么地准确。

我使用的代码可以在下面找到 ——

安装transformer包：

```
!git clone git@github.com:huggingface/transformers.git
!cd transformers
!pip install .
```

```
import scipy
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('bert-base-nli-mean-tokens')
```

语料如下：

```
# Get a sample corpus to search over
_c="""
Coronavirus:
White House organizing program to slash development time for coronavirus vacci
Trump says he is pushing FDA to approve emergency-use authorization for Gilead
AstraZeneca to make an experimental coronavirus vaccine developed by Oxford Un
Reopening:
Inconsistent patchwork of state, local and business decision-making on reopeni
White House risks backlash with coronavirus optimism if cases flare up again (
Florida plans to start reopening on Monday with restaurants and retail in most
California Governor Newsom plans to order closure of all state beaches and par
Japan preparing to extend coronavirus state of emergency, which is scheduled t
Policy/Stimulus:
Economists from a broad range of ideological backgrounds encouraging Congress
Global economy:
China's official PMIs mixed with beat from services and miss from manufacturin
China's Beige Book shows employment situation in Chinese factories worsened in
Japan's March factory output fell at the fastest pace in five months, while re
Eurozone economy contracts by 3.8% in Q1, the fastest decline on record (FT)
US-China:
Trump says China wants to him to lose his bid for re-election and notes he is
Senior White House official confident China will meet obligations under trad d
Oil:
Trump administration may announce plans as soon as today to offer loans to oil
Munchin says Trump administration could allow oil companies to store another s
Norway, Europe's biggest oil producer, joins international efforts to cut supp
IEA says coronavirus could drive 6% decline in global energy demand in 2020 (F
Corporate:
Microsoft reports strong results as shift to more activities online drives gro
Facebook revenue beats expectations and while ad revenue fell sharply in March
Tesla posts third straight quarterly profit while Musk rants on call about nee
eBay helped by online shopping surge though classifieds business hurt by closu
Royal Dutch Shell cuts dividend for first time since World War II and also sus
```

```
Chesapeake Energy preparing bankruptcy filing and has held discussions with le
Amazon accused by Trump administration of tolerating counterfeit sales, but co
Trump contradicts US intel, says Covid-19 started in Wuhan lab.
```

```
# Convert the corpus into a list of headlines
corpus=[i for i in _c.split('\n')if i != ''and len(i.split(' '))>=4]

# Get a vector for each headline (sentence) in the corpus
corpus_embeddings = model.encode(corpus)

# Define search queries and embed them to vectors as well
queries = [
    'The economy is more resilient and improving.', 'The economy is in a lot c
query_embeddings = model.encode(queries)

# For each search term return 5 closest sentences
closest_n = 5
for query, query_embedding in zip(queries, query_embeddings):
    distances = scipy.spatial.distance.cdist([query_embedding], corpus_embeddi

    results = zip(range(len(distances)), distances)
    results = sorted(results, key=lambda x: x[1])

    print("\n\n======================\n\n")
    print("Query:", query)
    print("\nTop 5 most similar sentences in corpus:")

    for idx, distance in results[0:closest_n]:
        print(corpus[idx].strip(), "(Score: %.4f)" % (1-distance))
```

结果如下：

```
======================



Query: The economy is more resilient and improving.


Top 5 most similar sentences in corpus:
```

```
Microsoft reports strong results as shift to more activities online drives gro
Facebook revenue beats expectations and while ad revenue fell sharply in March
Senior White House official confident China will meet obligations under trad d
Economists from a broad range of ideological backgrounds encouraging Congress 
White House risks backlash with coronavirus optimism if cases flare up again (


======================


Query: The economy is in a lot of trouble.

Top 5 most similar sentences in corpus:
Inconsistent patchwork of state, local and business decision-making on reopeni
eBay helped by online shopping surge though classifieds business hurt by closu
China's Beige Book shows employment situation in Chinese factories worsened in
Eurozone economy contracts by 3.8% in Q1, the fastest decline on record (FT) (
China's official PMIs mixed with beat from services and miss from manufacturin


======================


Query: Trump is hurting his own reelection chances.

Top 5 most similar sentences in corpus:
Trump contradicts US intel, says Covid-19 started in Wuhan lab. (Score: 0.7472
Amazon accused by Trump administration of tolerating counterfeit sales, but co
Trump says China wants to him to lose his bid for re-election and notes he is 
Inconsistent patchwork of state, local and business decision-making on reopeni
White House risks backlash with coronavirus optimism if cases flare up again (
```

上面的例子很简单，但是说明了语义搜索的一个重要方面。人类需要几分钟才能找到最相似的句子。它使我们能够在不需要人工参与的情况下在文本中查找特定信息，这意味着我们可以以计算机速度在成千上万个文档中搜索我们关心的短语。

这项技术已经被用来在两个文档中找到相似的句子。或者季度收益报告中的关键信息。例如，通过这种语义搜索，我们可以很容易地找到Twitter、Facebook、Snapchat等所有社交公司的日常活跃用户。尽管他们定义和叫法的是不同的——日活跃用户(DAU)或月活跃用户(MAU)或可盈利活跃用户

(mMAU)。由BERT支持的语义搜索可以发现所有这些表面形式在语义上意味着相同的东西 —— 一种性能的衡量，它能够从报告中提取我们感兴趣的句子。

对冲基金利用语义搜索来解析和展示季度报告(10-Q/10-K)中的指标，并在它们发布后立即将其作为量化交易信号，这不是一个遥远的想法。

上面的实验显示了语义搜索在过去的一年里取得了怎样的效果。

## 找到相似的句子 —— 聚类

使用这些句子向量嵌入的另一种主要方式是用于聚类。我们可以快速地将单个文档中的句子或多个文档中的句子聚成相似的组。

通过使用上面的代码，我们可以利用sklearn中的一个简单的k-means方法。

```
from sklearn.cluster import KMeans
import numpy as npnum_clusters = 10

clustering_model = KMeans(n_clusters=num_clusters)
clustering_model.fit(corpus_embeddings)
cluster_assignment = clustering_model.labels_

for i in range(10):
    print()
    print(f'Cluster {i + 1} contains:')
    clust_sent = np.where(cluster_assignment == i)
    for k in clust_sent[0]:
        print(f'- {corpus[k]}')
```

同样，对于一台机器来说，结果是准确的。这里有几个聚类 ——

```
Cluster 2 contains:
- AstraZeneca to make an experimental coronavirus vaccine developed by Oxford
- Trump says he is pushing FDA to approve emergency-use authorization for Gile

Cluster 3 contains:
- Chesapeake Energy preparing bankruptcy filing and has held discussions with
- Trump administration may announce plans as soon as today to offer loans to o
- Munchin says Trump administration could allow oil companies to store another

Cluster 4 contains:
```

```
- Trump says China wants to him to lose his bid for re-election and notes he i
- Amazon accused by Trump administration of tolerating counterfeit sales, but
- Trump contradicts US intel, says Covid-19 started in Wuhan lab. (The Hill)
```

## 总结

有趣的是，ElasticSeach现在有了dense向量的用法：https://www.elastic.co/blog/text- similar-search with-vectors-in-elasticsearch，可以和其他的工业界的快速比较两个向量的工具相比，如Facebook的faiss。这个技术是很尖端的，但具有很强的操作性，会在几周内推出。先进的人工智能触手可及，任何人都知道该寻找什么。

— END —

英文原文：https://towardsdatascience.com/cutting-edge-semantic-search-and-sentence-similarity-53380328c655