Document similarity

- James Singleton
- 30 September 2020

Aims

- document vectors can be created for text content in the content store, using pre-trained models
- these vectors can be used to establish similarity between documents
- using document vectors, a 2 dimensional (self-organised) map, called SOMs, can be created, representing the entire government corpus on a single page, for navigation of the corpus in relation to a search document

Progress in a nutshell:

- contents are first parsed from the content store into clean text, removing meta data and XML tags
- document vectors have been created using the Universal Sentence Encoder v4 (June 2020) from Google, and finding the average of the sentence vectors for a document
- a trained Approximate Nearest Neighbour tree was built for the curated text
- for new text, similar documents can be found in the Approximate Nearest Neighbour tree
- for existing text, (near) duplicate documents are simply found using the distance metrics
- separate to this piece of work, SOMs were created with heatmaps for similarity, to highlight distinct areas of the corpus which are related to a query

Explainer - Universal Sentence Encoder

The Universal Sentence Encoder compresses a document into a dense vector, of length 512. The reason for doing this is it compresses the meaning of a document into a uniform and compact shape. The Universal Sentence Encoder is a pretrained model from Google, trained on a larger corpus of documents. Using this model is transfering learning from other documents to our documents. Positions on the vector represent abstract concepts, which are unrelated to each other. The magnitude of values for an attribute reflect its strength.

Explainer - Cosine angle

Vectors have direction and magnitude (to quote Despicable Me). For any two documents, represented as vectors, their similarity to each other can in part be expressed by the geometric angle between them. Basically, if the document vectors point in the similar direction, they are similar. The cosine operation removes the magnitude and looks solely at the direction. Using the cosine formula, which is the product of two vectors divided by their magnitudes, we find the similarity between documents.

Explainer - Approximate Nearest Neighbours Oh Yeah

Approximate Nearest Neighbours solves the problem of speed of search for closely related terms. Naively trying to find the nearest neighbours for an item, on a case by case basis, would require repeatedly calculating the n-1 pairwise relationships, where n is the number of documents. To circumvent this problem, the corpus can be mapped as a tree, beforehand, on a one time basis. The tree is built by repeatedly bisecting the search space. Its a very simple idea. Each time the space is bisected, the number of elements in the resulting two spaces is halved. In this way, an address in hyperspace is created for an item. This index tree is then saved for subsequent use when a query is run.

Then, once an item is queried, the nearest neighbours calculation is preceded by finding the relevant address space in the index tree. Then looking in the space (and neighbouring spaces if requested), and only calculating the nearest neighbours in that space. This speeds up the search time, in proportion to 2**number of branches to the tree. This is an established problem which Spotify has addressed, but for music recommendations (they compress music into embedding vectors too). The Spotify library (ANNOY - Approximate Nearest Neighbours Oh Yeah) is the most user friendly, so that has been used here. Indexation of the space is initially time consuming, but after that is completed, the search times are the order of milliseconds. A search is performed by expressing a document as the embedding vector.

Example search and response

- Specify some search text (this could be a document someone is preparing for content).

```
test_text = [ 'Ministers are being accused of "sticking their fingers in their
ears" over the possibility that next year\'s public exams may have to be
cancelled.', 'The National Education Union says higher Covid-19 infection rates
and more pupils being sent home, makes holding exams unfair and less likely.',
'It argues that as the situation develops, using a system of centre assessed
grades may become inevitable.', 'The government is considering a slight delay
GCSE and A-levels next summer.', 'But NEU joint general secretary Mary Bousted
said this was a position that was becoming "increasingly untenable" and teachers
urgently needed to know what evidence of pupil achievement they needed to
collect so fair assessments could be made.', 'Her comments come a day after
national attendance figures revealed one in six secondary schools were not fully
open to all pupils last week, with 16% having to send at least some pupils home
to self-isolate amid a rise in virus cases.', 'It also comes after the vice-
chancellors of several universities called for next year\'s exams to be
cancelled, and for the focus to be on pupils catching up missed learning
instead.', 'Dr Bousted said national figures showed 200,000 children and young
people were not in school last week, and that with 7,000 new cases nationally
yesterday alone, disruption was inevitable.', '"All of that makes it more and
more difficult to see that students will get the opportunity to consistently be
in school across the country," she said.', '"As the situation develops it may
become inevitable that what we have to move to is a system of centre-assessed
grades... everybody appears to agree that this is a real possibility - that we
won\'t be able to do exams.', '"The only body which is sort of sticking its head
in the sand, sticking its fingers in its ears, is the government, and that is
what they have done consistently in this crisis.' ]
```

- Create an embedding vector for the query text:

```
embedding = document_embedding(test_text)
```

- Querying the ANN address ('u') list for related documents, using the embedding vector for the search document:

```
print(u.get_nns_by_vector(embedding, 4, include_distances=True))
```

- Which gives us the following list of documents and their distance from the query:

'([26042, 12673, 14202, 18615], [0.7543756365776062, 0.7550346851348877, 0.7634738087654114, 0.7896975874900818])'

- We look at the first document:

```
text_df['doc_text'][26042]
```

```
'The exam regulator, Ofqual, and exam boards will work with teachers to provide
grades to students whose exams have been cancelled this summer, following our
actions to slow the spread of coronavirus. University representatives have
confirmed that they expect universities to be flexible and do all they can to
support students and ensure they can progress to higher education. This year’s
summer exam series, including A levels, GCSEs and other qualifications, and all
primary assessments, have been cancelled as we fight to stop the spread of
coronavirus. The Government’s priority is now to ensure affected students can
move on as planned to the next stage of their lives, including going into
employment, starting university, college or sixth form courses, or an
apprenticeship in the autumn. This means ensuring GCSE, A and AS level students
are awarded a grade which fairly reflects the work that they have put in. There
will also be an option to sit an exam early in the next academic year for
students who wish to.\nOfqual will develop and set out a process that will
provide a calculated grade to each student which reflects their performance as
fairly as possible, and will work with the exam boards to ensure this is
consistently applied for all students. The exam boards will be asking teachers,
who know their students well, to submit their judgement about the grade that
they believe the student would have received if exams had gone ahead. To produce
this, teachers will take into account a range of evidence and data including
performance on mock exams and non-exam assessment – clear guidance on how to do
this fairly and robustly will be provided to schools and colleges. The exam
boards will then combine this information with other relevant data, including
prior attainment, and use this information to produce a calculated grade for
each student, which will be a best assessment of the work they have put in.
Ofqual and exam boards will be discussing with teachers’ representatives
before...
```