Terminology-based Text Embedding for Computing Document Similarities on Technical Content

Hamid Mirisae, Eric Gaussier, Cedric Lagnier, Agnes Guerraz

July 2019
Outline

• Introduction

• Related work

• Proposed method

• Experiments

• Conclusion
Businesses need to work with startups —> they need “good” info
crawl the web (startups), process them, provide structured info

- Concours I-LAB.
- Concours de l’innovation.
- Partnership with LIG.
How does it look like?

SHORT DESCRIPTION
Skopai develops an AI-based platform acting as a smart start-ups directory designed to propose a real-time, objective and complete knowledge of any start-up worldwide.

FOUNDERS
Skopai was founded in 2017 as a spin-off from the LIC and the Grenoble Alpes Data Institute by Agnès Guerraz (CEO), Bruno Sportisse and Éric Gaussier (Scientific Advisor).

INDUSTRY
It is a AI, PLATFORM and big data company focused on start-ups analysis.

PURPOSE
The ambition of Skopai is to help its customers to identify the start-ups that can interest them and to give them an up-to-date 360-degree view of these start-ups.

MARKET
Its market is Europe on business qualification and innovation expertise.

BUSINESS MODEL
Its business model is B2B.

REFERENCES
It has concluded partnerships with LIC (Grenoble Alpes Data Institute), BPI, and EY.
It seems to have concluded a partnership with EY following the EY Open Innovation contest to co-create an offer with the company, and to create a special offer for the Vivatech conference. It seems to already have concluded other contracts with customers.
Introduction
Introduction

• Finding relevant documents: on everyday basis.

• The principle is (almost) always the same:
  1. Define the space and the similarity measure.
  2. Take everything to that space.
  3. Find the closest documents in the space.

• Query ~ document:
  • closest papers to “Building Representative Composite Items” on arXiv?

How to define the space?
How to represent documents?
Related work (motivation?)

- Classic: tf-idf + cosine.
- KNN with tf-idf —> text classification [2014].
- Text representation: tf-idf, LSI and multiword —> text classification [2011].
- tf-idf is nice & helps in many tasks such as topic modelling, but…
- Let’s be more contextual:
  - Go to word level (word2vec).
  - Represent words such that they carry semantical features (based on co-occurrence).
  - $\text{vec(\text{King})} - \text{vec(\text{Man})} + \text{vec(\text{Queen})} \sim \text{vec(\text{Woman})}$
  - $\text{most_similar(car)} = [\text{cars, vehicle, automobile}]$
  - Many variations: doc2vec, sent2vec, combine with tf-idf, …
Motivation

Given a document, find similar documents in a “selective” fashion.

Do you *seriously* read the introduction and/or related work section of papers all the time?

**Focus on the important parts of the document.**

As mentioned previously, calculating document similarity and, consequently, finding similar documents is at the core of many ML tasks. Sometimes, however, focusing on the entire document may not lead to capturing desired similar documents as not all parts of a document have the same importance level. For instance, if a document describes a novel device for people suffering from diabetes, then taking the entire document may not necessarily result in the similar documents talking about the same particular issue, but rather about the medical domain in general. Note that this issue is different from

With that in mind, one can simply use the nodes of the main core to construct the keyphrases using different approaches, by for instance taking the words corresponding to the connected nodes. We use a slight modification of the $k$-core method in order to, first, extract the keyphrases of size 2, *i.e.* combination of two and only two terms, and, then, rank the sentences. Using those ranked sentences, we propose a technique to embed the document such that it encodes the main technical content of the document. The proposed approach is detailed in the following.
Proposed method: the big picture!

1. Extract keywords and/or keyphrases (composite keywords) of the document.

2. Score the sentences of the document based on the (composite) keywords they contain.

3. Pick a way to embed the sentences.

4. Embed the document as weighted average of the embeddings of its sentences.
Extracting (composite) keywords: use graphs!

Graph = Nodes + Edges

• Many problems can be formulated and/or interpreted via graph structure.

• In NLP:
  
  • Nodes —> entities (words, sentences, paragraphs, etc).
  
  • Edges —> relation with them (semantic, co-occurrence, etc).

• [Rousseau et al.] graph-of-words:
  
  • Nodes —> terms of the documents.
  
  • Edges —> if two terms co-occur in a fixed-size window.
The proposed method can be used to find similar documents, particularly when the technical content is concerned for finding relevant documents.

proposed method similar documents technical content relevant documents

- proposed method similar —> \{proposed, method\}, \{proposed, similar\}, \{method, similar\}
- method similar documents —> \{method, similar\}, \{method, document\}, \{similar, document\}
- similar documents technical —> …
- documents technical content —> …
- technical content relevant —> …
- content relevant documents —> …

(un)weighted (un)directed
$\mathcal{H}_k = (\mathcal{V}', \mathcal{E}')$ is called a k-core or a core of order $k$ of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ iff $\mathcal{E}' \subseteq \mathcal{E}$, $\mathcal{V}' \subseteq \mathcal{V}$ and $\forall \nu \in \mathcal{V}'$, $\text{Deg}(\nu) \geq k$, and $\mathcal{H}_k$ is the maximal graph with such property.

- Core with max $K$ $\rightarrow$ **main core**
- **Idea**: it’s important to be central, but your neighbours are also important!
- [Rousseau & Vazirgiannis]:
  - Main core $\rightarrow$ **keywords & keyphrases**
  - Better than HITS and PageRank.
  - No hyperparameter.
TDE: Terminology-based Document Embedding (informally)

We develop artificial pancreas which acts like an insulin pump. (score = 23.2 + 19.4 = 42.6)

Via a clinical test, we evaluated our insulin pump. (score = 19.4 + 14.6 = 34)

Do the math!

Score the sentences based on their keyphrases

- artificial pancreas (23.2)
- insulin pump (19.4)
- clinical test (14.6)

Score them based on their core & their edge weight
TDE: Terminology-based Document Embedding (formally)

\[ \mathcal{C} = \{c_1, \cdots, c_k\} \]
\[ \mathcal{T}_{c_i} = \{(t, t') | t \in c_i \land t' \in c_i\} \]
\[ \vec{d} = \frac{1}{\sum_{s \in S} \Gamma(s)} \sum_{s \in S} \vec{s} \times \Gamma(s) \]
\[ \Gamma(s) = \sum_{i=1}^{i=k} \sum_{(t, t') \in \mathcal{T}_{c_i}} \phi((t, t')) \]
\[ \phi((t, t') \in \mathcal{T}_{c_i}) = Deg(t, t') \times F(c_i) \]
\[ F(c_i) = (k - i + 1)^{-1} \]

**Algorithm 1** Terminology-based Document Embedding

**Input:** Set \( S \) containing all sentences of document \( d \), \( \mathcal{T}_{c_i} (1 \leq i \leq k) \): keyphrases of each core

**Output:** \( \vec{d} \): the embedding of \( d \)

1: \( w = 0 \)
2: \( \vec{d} = \vec{0} \)
3: **for all** \( s \in S \) **do**
4: \( w_s = 0 \)
5: **for** \( i = 1 \) **to** \( k \) **do**
6: **for all** \( (t, t') \in \mathcal{T}_{c_i} \) **do**
7: **if** \( (t, t') \in s \) **then**
8: \( w_s = w_s + \frac{Deg(t, t')}{i} \)
9: **end if**
10: **end for**
11: **end for**
12: \( \vec{d} = \vec{d} + (w_s \times \vec{s}) \) \( \backslash \vec{s} \) is the embedding of \( s \)
13: \( w = w + w_s \)
14: **end for**
15: \( \vec{d} = \frac{\vec{d}}{w} \)
16: **RETURN** \( \vec{d} \)
Experiments

• Baselines:
  - doc2vec: directly embed a document.
  - TWA: tf-idf weighted average of words of the document.
  - TDE: how to represent a sentence?
    - sent2vec: learn a model to embed sentences.  \( TDE_{s2v} \)
    - (tf-)idf weighted average of it’s words. \( TDE_{iw} \)
Experiments: dataset

- Crawling websites of 68K startups (3.4M pages).
- Filter non-English, take pages with texts —> 43K startups with 2.8M pages.
- Document = combination of some pages of the startup.

Terms of use

Welcome to SKOPAI platform!

This platform is only accessible to specific users.

Access to, and use of the SKOPAI platform available at the address www.skopai.com (hereafter referred to as the "Website") and the services it proposes (hereafter referred to as the "Services") are subject to compliance with these general conditions.

These conditions apply as from 23/01/2019.
Experiments: training, evaluation & results

- **100 documents**, four **domains**: {medical, agriculture, energy, biology}, **scores** {1, ..., 5}.

- Metric: Normalized Discounted Cumulative Gain, **NDCG@1**, **NDCG@5**.

\[
DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i + 1)}\]
\[
NDCG_p = \frac{DCG_p}{IDCG_p}
\]

- Train the doc2vec on the dataset.

- Train the sent2vec on all the sentences (omitting the stopwords) of the dataset.

<table>
<thead>
<tr>
<th></th>
<th>D2V</th>
<th>TWA</th>
<th>TDE_{i,w}</th>
<th>TDE_{s2v}</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@1</td>
<td>0.26</td>
<td>0.54</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>NDCG@5</td>
<td>0.24</td>
<td>0.60</td>
<td>0.60</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Experiments: more details

HERE

sentences —> only English, no stopwords.

IN PRACTICE

- We use $TDE_{iw}$
- Multilingual parallel w2vec models (EN, FR, DE, ES).
- Only nouns & adj.
- Check if the keyphrase is actually a keyphrase.
  - language specific RE checking (NN-NN, NN-ADJ, …)
  - imprime 3D, 3D printer, satellite de communication, …

The proposed method can be used to find similar documents, particularly when the technical content is concerned for finding relevant documents.

- proposed method similar
- method similar documents
- similar documents technical
- documents technical content
- technical content relevant
- content relevant documents
Experiments: examples

SKOPAI (top-3 sentences)

We have a strong experience of the innovation ecosystem and how it works: research and technology transfer in academia, R&D in tech or industry corporations, venture capital or government innovation policy.

Nous construisons une plate-forme de référence pour la technologie, fournissant en temps réel une connaissance complète sur toute startup dans le monde entier.

Startup assessment depends on the quality and context of the person performing it – for example chief innovation officer, product managers, R&D engineers, investors, buyers, legals, etc.

SKOPAI (top-3 similars)

We build on a longstanding experience in corporate, startup, not-for-profit and public service organizations. We stand for collaboration that allows businesses to thrive. This is why we focus on enabling alliances that foster innovation and redesign business models.

We are Building the only all-in-one innovation & startup ecosystem platform for the cloud power future.

Startup assessment depends on the quality and context of the person performing it – for example chief innovation officer, product managers, R&D engineers, investors, buyers, legals, etc.
Conclusions

• Use graph-based methods to extract similar documents.

• The general framework is valid for any sort of sentence embedding.

• Focus on the technical content (via keyphrases).

• Outperform the state-of-the-art in terms of NDCG.

• Could be also used to rank sentences.
We are hiring...

https://www.skopai.com/join-us/

Full Stack Engineer
Data Scientist