



# **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?

The screenshot displays the Skopai website interface. At the top, a teal navigation bar contains the Skopai logo (a stylized globe) and the text 'Skopai Deep Tech Insights'. To the right of the logo are links for 'WHAT'S NEW?', 'MY STARTUPS', 'CONTACT US', and 'ADMIN'. Further right, a user profile for 'hamid.mirisaee' is visible. Below the navigation bar, a dark teal banner features the company name 'SKOPAI' and a series of menu items: 'AT A GLANCE', 'TEAM & NETWORK', 'MARKET & COMPETITION', 'PRODUCT & TECHNOLOGY', 'FINANCE', and 'INSIGHT'. The main content area is divided into several sections: a 'SHORT DESCRIPTION' section with a small globe icon and text stating 'Skopai develops an AI-based platform acting as a smart start-ups directory...'; a 'FOUNDERS' section detailing the company's origin in 2017; an 'INDUSTRY' section describing it as an AI and PLATFORM company; a 'PURPOSE' section outlining the goal of helping customers identify start-ups; a 'PRODUCT' section (partially visible); a 'MARKET' section identifying the market as Europe; a 'BUSINESS MODEL' section stating it is B2B; and a 'REFERENCES' section mentioning partnerships with LIG, BPI, and EY. On the left side, a sidebar provides key information: 'STAGE' (Early Stage), 'COUNTRY' (France), 'CREATION' (July 21, 2017), 'SIZE' (from 11 to 50 employees), 'TAGS' (Investment, Machine Learning, Bigdata, Platform, AI), 'WEB SITE' (https://www.skopai.com/), and 'CONTACT' (contact@skopai.com).

← Skopai Deep Tech Insights

WHAT'S NEW ? | MY STARTUPS | CONTACT US | ADMIN

HA hamid.mirisaee

SKOPAI

AT A GLANCE | TEAM & NETWORK | MARKET & COMPETITION | PRODUCT & TECHNOLOGY | FINANCE | **INSIGHT**

**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 LIG and the Grenoble Alpes Data Institute by Agnès Guerraz (CEO), Bruno Sportisse and Éric Gaussier (Scientific Advisor).

**INDUSTRY**

It is a AI and PLATFORM 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.

**PRODUCT**

Skopai develops an AI-based platform acting as a smart 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 LIG (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 Vivatch conference. It seems to already have concluded other contracts with customers.

**STAGE** ⓘ Early Stage

**COUNTRY** France

**CREATION** July 21, 2017

**SIZE** from 11 to 50 employees

**TAGS**

INVESTMENT

MACHINE LEARNING

BIGDATA PLATFORM AI

**WEB SITE** <https://www.skopai.com/>

**CONTACT** [contact@skopai.com](mailto:contact@skopai.com)

# Introduction

Google Scholar

graph based document similarity

Articles About 450,000 results (0.11 sec)

My profile My library

Any time  
Since 2019  
Since 2018  
Since 2015  
Custom range...

Sort by relevance  
Sort by date

include patents  
 include citations

Create alert

[PDF] **Graph-based ranking algorithms for sentence extraction, applied to text summarization** [\[PDF\] aclweb.org](#)  
R Mihalcea - Proceedings of the ACL Interactive Poster and ..., 2004 - aclweb.org  
... but rather it takes into account information re- cursively drawn from the entire text (**graph**). Through the **graphs** it builds on texts, TextRank identifies con- nections between various entities in ... text units, and the strength of the recommendation is recursively computed **based** on the ...  
☆ [Cited by 415](#) [Related articles](#) [All 14 versions](#)

**Phrase-based document similarity based on an index graph model** [\[PDF\] semanticscholar.org](#)  
KM Hammouda, [MS Kamel](#) - 2002 IEEE International ..., 2002 - ieeexplore.ieee.org  
**Document** clustering techniques mostly rely on single term analysis of the **document** data set, such as the vector space model. To better capture the structure of **documents**, the underlying data model should be able to represent the phrases in the **document** as well as ...  
☆ [Cited by 81](#) [Related articles](#) [All 9 versions](#)

**Efficient phrase-based document similarity for clustering**  
H Chim, [X Deng](#) - IEEE Transactions on Knowledge and Data ..., 2008 - ieeexplore.ieee.org  
... As a result, other five base clusters in the **graph** will not form a single cluster in the ... the STD model is trying to keep the sequential order of each word in a **document**, the same ... In our phrase-**based document similarity**, the word term is replaced by the node term in the suffix tree ...  
☆ [Cited by 196](#) [Related articles](#) [All 9 versions](#)

**Efficient graph-based document similarity** [\[PDF\] semanticscholar.org](#)  
C Paul, [A Rettinger](#), [A Mogadala](#), [CA Knoblock](#)... - European Semantic ..., 2016 - Springer  
Assessing the relatedness of **documents** is at the core of many applications such as **document** retrieval and recommendation. Most **similarity** approaches operate on word- distribution-**based document** representations-fast to compute, but problematic when ...  
☆ [Cited by 31](#) [Related articles](#) [All 5 versions](#)

**Learning a concept-based document similarity measure** [\[PDF\] researchgateway.ac.nz](#)  
I Huang, D Milne, F Frank ... - Journal of the American ..., 2012 - Wiley Online Library

# 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  $\rightarrow$  text classification [2014].
- Text representation: tf-idf, LSI and multiword  $\rightarrow$  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).
  - $\text{vec}(\text{King}) - \text{vec}(\text{Man}) + \text{vec}(\text{Queen}) \sim \text{vec}(\text{Woman})$
  - $\text{most\_similar}(\text{car}) = [\text{cars}, \text{vehicle}, \text{automobile}]$
  - 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  $\rightarrow$  entities (words, sentences, paragraphs, etc).
  - Edges  $\rightarrow$  relation with them (semantic, co-occurrence, etc).
- [Rousseau et al.] graph-of-words:
  - Nodes  $\rightarrow$  terms of the documents.
  - Edges  $\rightarrow$  if two terms co-occur in a fixed-size window.

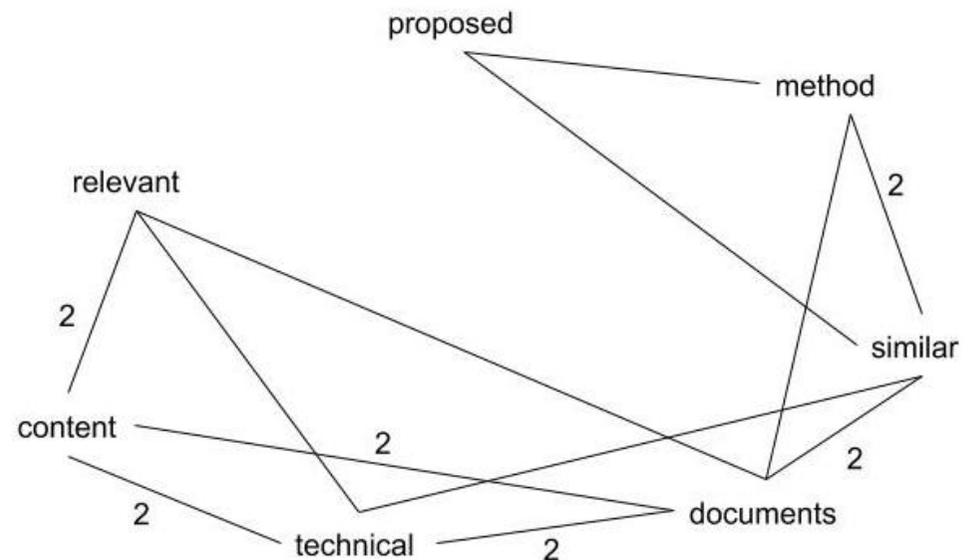
# Graph-of-words

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  $\rightarrow$  {proposed, method}, {proposed, similar}, {method, similar}
- method similar documents  $\rightarrow$  {method, similar}, {method, document}, {similar, document}
- similar documents technical  $\rightarrow$  ...
- documents technical content  $\rightarrow$  ...
- technical content relevant  $\rightarrow$  ...
- content relevant documents  $\rightarrow$  ...

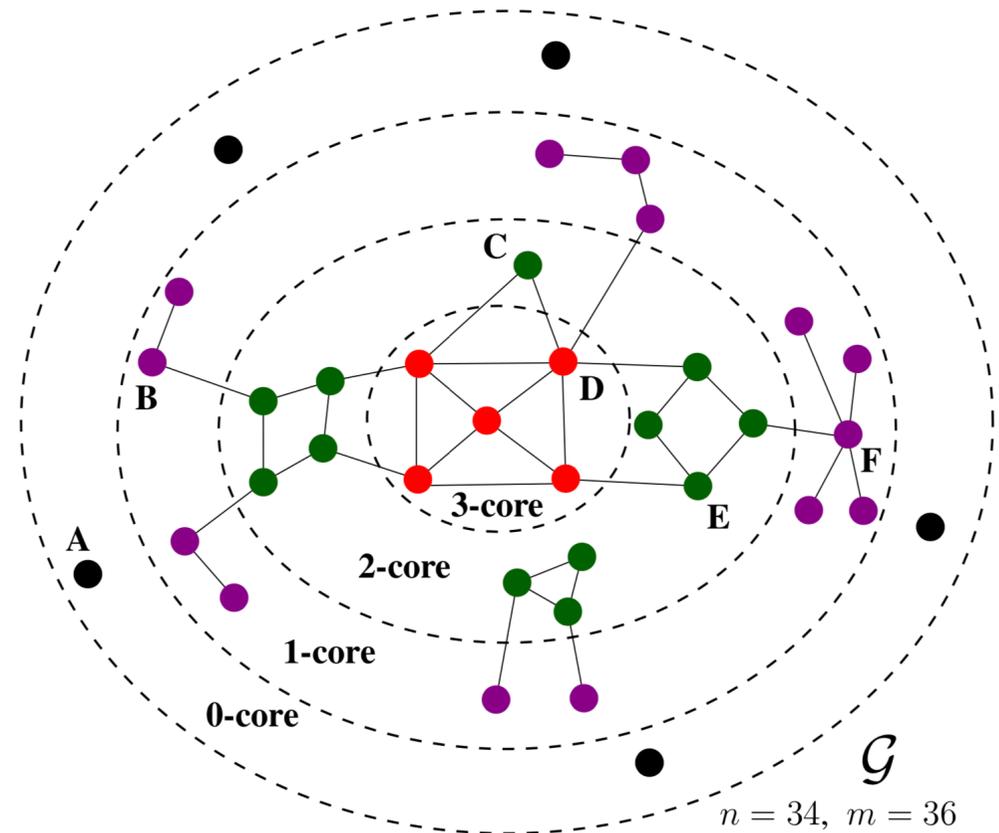
**(un)weighted**  
**(un)directed**



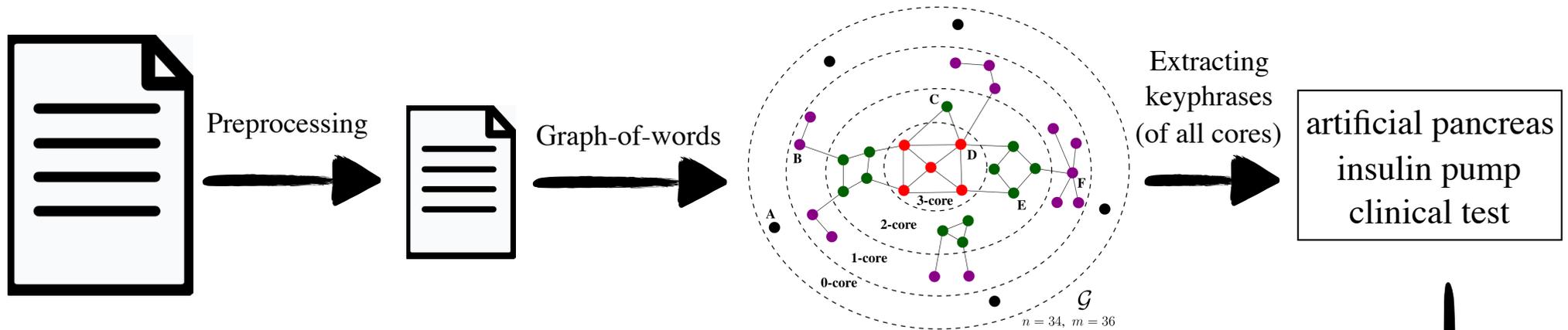
# K-core

$\mathcal{H}_k = (\mathcal{V}', \mathcal{E}')$  is called a  $k$ -core or a core of order  $k$  of  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  iff  $\mathcal{E}' \subset \mathcal{E}$ ,  $\mathcal{V}' \subset \mathcal{V}$  and  $\forall v \in \mathcal{V}'$ ,  $Deg(v) \geq k$ , 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)



$$\vec{d} = \frac{42.6}{(42.6 + 34)} \times \vec{s}_1 + \frac{34}{(42.6 + 34)} \times \vec{s}_2 = 0.55 \times \vec{s}_1 + 0.45 \vec{s}_2$$

Do the math!

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)

Score them based on their core & their edge weight

Score the sentences based on their keyphrases

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

# TDE: Terminology-based Document Embedding (formally)

$$\mathcal{C} = \{c_1, \dots, c_k\}$$

$$\mathcal{T}_{c_i} = \{(t, t') | t \in c_i \wedge t' \in c_i\}$$

$$\vec{d} = \frac{1}{\sum_{s \in S} \Gamma(s)} \sum_{s \in S} \vec{s} \times \Gamma(s)$$

$$\Gamma(s) = \sum_{i=1}^{i=k} \sum_{\substack{(t, t') \in \mathcal{T}_{c_i} \\ (t, t') \in s}} \phi((t, t'))$$

$$\phi((t, t') \in \mathcal{T}_{c_i}) = Deg(t, t') \times F(c_i)$$

$$F(c_i) = (k - i + 1)^{-1}$$

**ALL CORES  
ONLY KEYPHRASES  
ONLY 2-WORD**

---

## 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:   $\vec{d} = \vec{d} + (w_s \times \vec{s})$      $\backslash\backslash$   $\vec{s}$  is the embedding of  $s$ 
13:   $w = w + w_s$ 
14: end for
15:  $\vec{d} = \frac{\vec{d}}{w}$ 
16: RETURN  $\vec{d}$ 

```

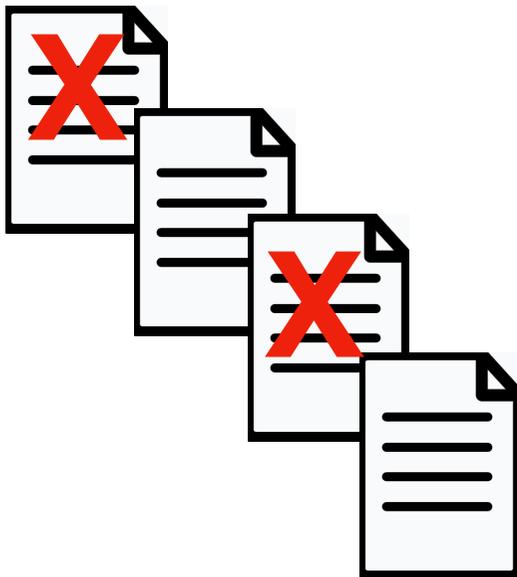
---

# 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.



**Skopai**  
Deep Tech Insights

## 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](http://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.

	D2V	TWA	TDE <sub>iw</sub>	TDE <sub>s2v</sub>
NDCG@1	0.26	0.54	0.63	<b>0.69</b>
NDCG@5	0.24	0.60	0.60	<b>0.65</b>

# Experiments: more details

## HERE

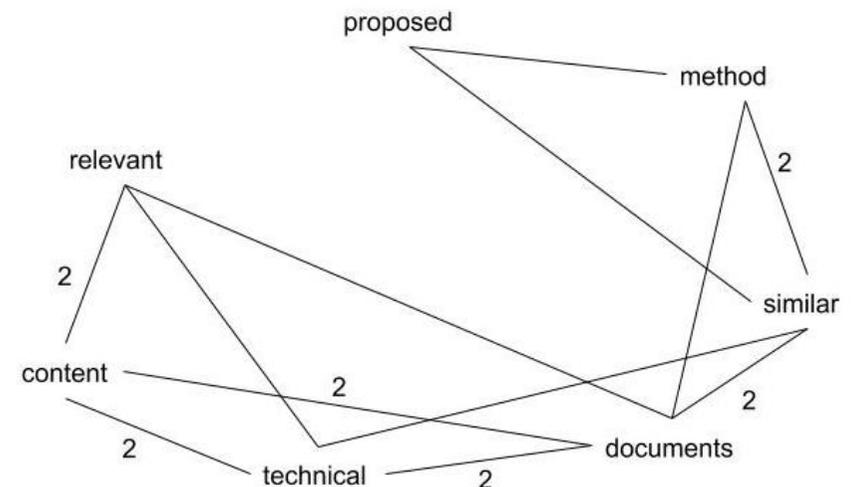
sentences → only English, no stopwords.

## 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, ...)
  - imprimente 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 & start-up 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.



Thank you!

**We are hiring...**

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

Full Stack Engineer  
Data Scientist