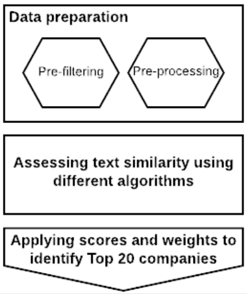
Expanding Siemens’ AI Healthcare Ecosystem

Team HacKings for the WHU Hackathon

On the task:

To cope with an ever-increasing need for innovative solutions, large enterprises recently started to foster ecosystems of cutting edge startups. In their mission to disrupt the healthcare sector, Siemens requires expertise in the field of artificial intelligence (AI). Finding the most suitable partners from countless AI startups imposes a crucial task on Siemens. The following outline describes the use of machine learning algorithms to select ventures based primarily on the similarity of company descriptions. By design, the step-by-step guide summarizes our approach an the main consideration in the process of aiding Siemens.



Data preparation

Pre-filtering

Pre-processing

Assessing text similarity using different algorithms

Applying scores and weights to identify Top 20 companies

Data preparation:

The raw data set, which was retrieved from Crunchbase, includes key information on close to 20,000 companies. Before applying any algorithm, we first narrowed down the solution space by pre-filtering the company list for relevant categories. To do so, keywords taken from Siemens’ description were used as a filter for the company categories, which reduced the number of thoroughly analyzed companies to approx. 2,000. We decided not to filter for specific geographies as we found that Siemens’ AI efforts in healthcare are diversified globally and do not concentrate on one isolated geography. Additionally, in the standard course of pre-processing (lemmatization, punctuation, whitespace, stopword removal), the company descriptions were cleaned homogenously before feeding them into any algorithm to potentially further improve the results.

Keywords:

Health

Hospital

Diagnostics

Pharmaceutical

Medical

Care

Genetics

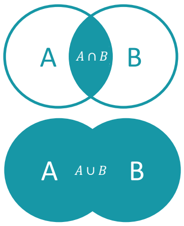
Biotechnology

Methods for assessing text similarity:

In the context of language processing, the **Jaccard** and **Cosine** similarities mark two fudamentally different technqiues to calculate the similarity between two or more documents. Due to its intuitiveness, the Jaccard similarity is a sensible introduction to the partner selection. In this approach, sentences are split into words and for each comparison, the intersect and union of different strings are compared. The higher the overlap relative to the total number of words, the higher the similarity measure. However, this approach does not consider the context of words. Consequently, more advanced algorithms have been employed to test the robustness of results. First, we applied **TFIDF**, a word-document mapping algorithm that evaluates how important a word is in a given document. Precisely, TFIDF is a measure of term frequency times inverse document frequency, where the latter is introduced to diminish the weight of terms that occur very frequently. Not only is TFIDF relatively easy to compute, but it also provides an intuitive measure of similarity, i.e. relative term frequency. However, it relies on a bag-of-words model and therefore does not capture positioning of words and semantics and is frequency-based, meaning it assumes the count of different words provides sufficient evidence for similarity between different strings. We also considered the possibility of using **GloVe**, an unsupervised method using dense vector representations, compared to TFIDF’s sparse vectors, however, we assessed the fact that GloVe’s count-based approach would not be a significant improvement compared to TFIDF. Hence, we went further and applied a neural-network-based document embedding procedure, **Doc2Vec**, to the data set. Not only does this approach pose an improvement on the prior in terms of taking word order into account, but such a model is also capable of learning from unlabeled data, making it a more appropriate alternative in our case. However, a Doc2Vec model still provides results which are context independent and usually requires to be trained on an extremely large corpus to yield meaningful results. Finally, to account for the drawbacks of the previous models, we applied the state-of-the-art **BERT** algorithm to the company descriptions. BERT makes use of a transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. As opposed to other directional models, e.g. **ELMo**, which read the text input sequentially (left-to-right or right-to-left), the BERT transformer encoder reads the entire sequence of words at once.

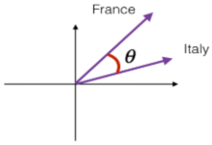
Jaccard similarity:

$$J(A,B) = \frac{A \cap B}{A \cup B}$$



Cosine similarity:

$$C(U,V) = \frac{\vec{U} * \vec{V}}{\|U\|x\|V\|}$$



France and Italy are quite similar

θ is close to 0°

$\cos(\theta) \approx 1$

TFIDF

Doc2Vec

BERT

GloVe

ELMo

Fuzzy

Scoring methodology and results:

Given the top 20 similar companies for each method, we assigned points based on a company‘s rank, i.e. Rank 1 receives 20 points and Rank 20 one point. Moreover, we applied different weightings to each method to calculate the final top 20 companies. BERT (50%) received the greatest weight given its novelty and complexity. Similarly, TFIDF (10%) and Doc2Vec (15%) received smaller weights to account for decreased complexity and other drawbacks described above. Lastly, we assigned the Jaccard Similarity (25%) the second-highest weight to gauge the different dimension it provides in terms of assessing similarities. We also acknowledge that such a ranking based on text similarities only provide an indication and are very much subject to other factors.

Top 5 companies:

Informatics In Context

Amplion

MedCircuit

surgical.ai

Zebra Medical Vision