Resume Shortlisting Automation

Review on approaches

The Context

Talent Acquisition leaders face a hard time selecting the right person while hiring. Our AI team at Infogen-labs is building a resume automation software to help Talent Acquisition team to select the right match for the position based on their skills, location and project experiences.
With the fast growth of Internet-based recruiting, there are a great number of CVs among hiring systems. To gain more attention from the recruiters, most resumes are written in diverse formats, including varying font size, font color, and tables. However, the diversity of format is harmful to data extraction and processing.

Approach 1: Regular Expression

Regular expressions were used to extract: Name, Location, Email Id, Experience, Phone Number. Resumes were classified by the frequency of keyword appearing.
Challenges faced by Model:
Extraction of Candidate name and location is difficult by Regular Expression having a wide variety of resume. Key-words in job description and skills in the resume are not exactly the same. This problem was taken care by string cosine similarity and word2Vec model.

Approach 2: How to find semantic similarity between two documents

This Doc2vec (aka paragraph2vec, aka sentence embeddings) modifies the word2vec algorithm to unsupervised learning of continuous representations for larger blocks of text, such as sentences, paragraphs or entire documents. As it is a Deep Learning model it requires a huge amount of data. With a limited amount of data, Doc2vec model doesn't give good accuracy.

We trained a model on more than 2000 resume from a different domain and we didn't achieve good accuracy.

Approach 3: Named Entity Recognition (default model)

Named Entity Recognition is a process where an algorithm takes a string of text (sentence or paragraph) as input and identifies relevant nouns (a variety of named and numeric entities, including companies, locations, organizations, and products) that are mentioned in that string. Default Named Entity Recognition model of spaCy gives low accuracy on resumes because resumes lag in context.
We tried to use Regex on NER results which improved accuracy.

Approach 4: Topic Modeling

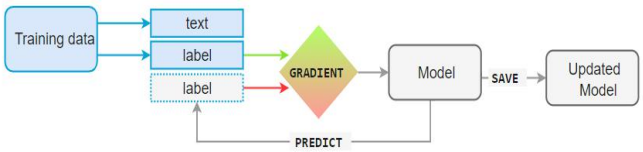
Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of a topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.
Resumes are summarized version of Documents. Hence Topic modeling doesn't give a good representation of the document. We have tried LDA, LSI, and HDP. LDA performs better amongst all.

Approach 5: NER spaCy training

SpaCy models are statistical and every "decision" they make – for example, which part-of-speech tag to assign, or whether a word is a named entity – is a prediction.
This prediction is based on the examples the model has seen during training. To train a model, you first need training data – examples of text, and the labels you want

the model to predict. This could be a part-of-speech tag, a named entity or any other information.

The model is then shown the unlabelled text and will make a prediction. Because we know the correct answer, we can give the model feedback on its prediction in the form of an error gradient of the loss function that calculates the difference between the training example and the expected output. The greater the difference, the more significant the gradient and the updates to our model.



We have trained model on Datatruks resume dataset https://www.kaggle.com/dataturks/resume-entities-for-ner We could achieve almost 99% accuracy on their test data.

Now we annotating our Resume database for Spacy NER training.

References:

[1] https://www.hindawi.com/journals/mpe/2018/5761287/

[2] http://www.ijircce.com/upload/2016/april/218_Intelligent.pdf

[3] https://www.researchgate.net/publication/221614548_PROSPECT_A_system_for_screening_candidates_for_recruitment

[4] https://openproceedings.org/2013/conf/edbt/MehtaPSVV13.pdf

[5] http://www.ierjournal.org/pupload/vol2iss7/Resume%20Parsing%20And%20Processing%20Using%20Hadoop.pdf

[6] https://www.ijert.org/research/proposed-system-for-resume-analytics-IJERTV5IS110274.pdf

[7] https://spacy.io/usage/training

[8] https://radimrehurek.com/gensim/

[9] https://rare-technologies.com/doc2vec-tutorial/