**Part 1 : Clustering**

In order to cluster the companies of the dataset, I wanted to use the variable TAGLINE which is the description of the company to get some keywords that describe best the company. I am hoping to get words that will be related to at least one of the four entities. I used a keyword extractor algorithm that extracts the noun phrase as keywords. Afterward, I put everything in lowercase, lemmatized the words and remove all the duplicates to get the final list of keywords. I also cleaned the TAGS to turn them into keywords using similar processes.

Secondly, I needed keywords to describe the four entities. I first selected some words that describe correctly, according to me, the entities but I wanted more words/synonyms. I used the wordnet tool to get the synonyms of the words I first chose and refine the list of keywords.

Then, I created the variable MAX\_OVERLAP that equals 0 if the overlap between TAGS and TAGLINE vs the entity keywords is bigger for the startups, 1 for the mature companies, 2 for universities and schools and 3 for the governmental or non-profit companies.

To continue, I wanted to use the fact that startups have to be created after 1990 and mature companies before. Thus, I cleaned the variable LAUNCH DATE in order to get a dummy variable (0 = after 1990 and 1 = before 1990). I classified the companies with MAX\_OVERLAP equals to 1 (mature companies) but with a launch date after 1990 as startups and vice versa. For the other companies, I kept the MAX\_OVERLAP result.

**Comments on part 1 :** I think the entity keywords need to be refined. One could use the dataset to see how each entity is described to improve the actual list. Also, using the hyponyms and hypernyms of wordnet could help. Moreover, some variables have not been used such as the name or growth stage that might be useful. One could also scrap the website to get more text to analyze. Algorithms/Technics such as Word2Vec or K-means should be investigated.

**Part 2 : Scrapping**

I scraped the website using selenium and beautifulSoup packages. I choose to extract the name of the company, legend, description, link of the website, year of creation, team size and location. For time reasons (I have exams next week), I chose to not extract the name and description of the founders.