

## Exercise 2

We decided to work with the dictionary *Positive* from the Loughran-McDonald set.

### 2.2. Explore whether the scores differ according to the meta data fields you gathered:

#### Without `tf_idf` weighting:

- President: Usually, each president has similar scores and appears in similar positions in the ranking (e.g. Carter Speeches have the highest scores). This might be because each president, like us, tends to use the same or similar words in all the speeches and therefore the score is similar.
- Year: It seems that the president plays a more important role than the year.
- Sources: Written speeches tend to have higher scores than Oral speeches.

**With `tf_idf` weighting:** On the other hand, when applying `tf_idf` weighting, the scores seem to be less dependent to the different metadata fields (e.g., the speeches of the same president are not as close in the ranking as with the previous method).

### 2.3 Do the answers to the previous question depend on whether tf-idf weighting is applied or not? Why do you think there is a difference in your answers?

Yes, the answers to the previous question are only applicable to when tf-idf is not applied.

When applying tf-idf, we take into account the frequency of the words in our different documents. So, if a word is giving a lot of score to a president that uses it frequently, when applying tf-idf the score of this word is going to be penalized because it is very frequent in our documents.