## Sentiment classification of company transcripts

Serena Xiao, Tuhin Harit

**Motivation:**

Berekshin et al (2017), Graham et al (2017), Biggerstaff et al (2015), Davidson et al (2015), Fioerdelisi et al (2014), among many others have demonstrated significant impact of corporate culture on the market, firm policies and performance. Most such papers use a combination of surveys, financial statement proxies (e.g. R&D for innovation), and external database proxies (e.g. patents for innovation). Recently papers such as Li et al (2018), and Loughran et al (2014) are using textual analysis to read company sentiments.

We want to use machine learning to generate sentiments score from company transcripts using similar methodology as Li et al (2018). Further we want to see find the relation of our innovation score with certain variables such as R&D expenditure, leverage etc.

For the purpose of this project we want to generate ‘green innovation’ and ‘racial sensitivity’ scores (these sentiments were not specifically addresses in Li et al (2018), but can be arguably thought of as a subset of ‘innovation’ and ‘respect’ values). Such scores might have a causal effect on firm’s market factors such as P/E, policies, and/or performance, such as R&D investment.

**Methods**

Li et al (2018) use Word2vec to generate sentiment score covering five cultural values, namely innovation, integrity, quality, respect and team-work. The authors borrow these values from Guiso et al 2015. The algorithm requires seed words for the sentiment, and uses that in a skip-gram model to calculate word-embedding vectors. Conceptually it follows Harris (1954) in that words tend to co-occur with neighboring words with similar meanings. Nearness is measured using cosine similarity. The paper borrows seed words from Guiso et al (2015). We plan to use “most frequently used words” algorithm as proposed by Loughran and Mcdonald (2014). They suggest using textual analysis to generate most frequently used words from some authoritative text on a sentiment (For e.g. a representative text on green innovation is “Accelerating Green Innovation”, by Michael Migendt). And after weeding out Stopwords and numbers, we can get the requisite seed words. The sentiments we focus on are “Green Innovation” and “Racial sensitivity” which can be arguably be thought of as a sub-category under innovation and respect.

1. Generate Seed words for Green innovation and Racial Sensitivity as the frequently used words in representative text on these two topics (and using codes from Laughran and McDonald)
2. Use these seed words, training dataset (consisting of text of randomly selected transcripts) and codes from Li et al, generate the dictionary for Green innovation and Racial Sensitivity
3. Give Green innovation and Racial Sensitivity score to each transcript (use codes from Laughran and McDonald)
4. Do prediction of value score on various variables

**Data**

1. Representative text books and research papers can be searched on google. Use freely available pdf files could be used. For Green Innovation I already have put some in the share folder
2. L&M code can be found here: <https://sraf.nd.edu/>
3. Transcripts can be obtained from wrds using python. I have the code and have shared in common folder
4. Li et al code can be found here: <https://github.com/MS20190155/Measuring-Corporate-Culture-Using-Machine-Learning>

We plan to use the text from company transcripts as available on WRDS. We plan to use the same methodology as Li et al 2018, using word2vec to train the model. The advantage of their method is that it is able to identify common phrases and treat them as words. Secondly, using cosine similarity ensures that a word is added in the dictionary for the value with which it has highest similarity. This is important for words which are synonyms for multiple cultural values. Also the vector representation algorithm is able to account for the context in which a word is used majority of times.

In addition to the papers already mentioned, we will read Textual analysis in accounting and finance: A survey by Loughran & McDonald (2016) , Erel, Stern & Weisbach (2018) and any other relevant text on the subject.