Metis Natural Language Processing Project – Sept 2021

Sam Reiff

Abstract

It's heuristic knowledge among (successful) stock pickers that there are usually 3 major factors that drive a stock's valuation, such as product releases, outlook for the business cycle, competitive positioning, etc. Identifying these drivers can be time consuming, especially for an unfamiliar company or industry. Using transcribed company/investor commentary, NLP can be a viable tool in understanding key topics of interest for individual documents and many documents over time, requiring a fraction of the effort compared to manually reading many of documents. This analysis serves as a proof of concept that there is value in building an NLP pipeline for company transcripts, particularly for industry analysts and asset managers.

Design

This analysis seeks to produce a meaningful output of topics and topic words as a demonstration of NLP value as applied to Texas Instruments' (ticker TXN) publicly available conference call transcripts, which are typically structured into two main components: **1) Prepared remarks:** scripted C-suite commentary delivered as a high-level overview of the current state of the company and industry; and **2) Q&A:** questions delivered by Wall Street analysts and investors ad-hoc, with unscripted responses by company management. Relevance of topics and topic words are assessed based on my domain knowledge of the semiconductor industry and knowledge of TXN from an analyst's point of view.

Data

This analysis leverages over 70 distinct documents that I bifurcated into prepared remarks and Q&A, resulting in ~150 total documents after text preprocessing; in total, the documents contained over 3 million individual words prior to cleaning. Data was acquired via Seeking Alpha's API. Naturally, the word content of these documents centered around financial terms (eg, 'free cash flow', 'gross margin', 'balance sheet'), though there was ample technology and industry terminology to draw meaningful topics regarding industry trends (eg, 'channel inventory', 'fab (fabrication facility)', 'embedded processing').

Text preprocessing was requisite for meaningful analysis. As a transcript, the names of management, analyst names, and boilerplate safe harbor statements were overwhelmingly numerous prior to cleaning, and were removed from the corpus in my analysis pipeline.

Algorithms/tools

- Python Pandas and NLTK for data pre-processing
- Python NLTK, Gensim for model implementation and analysis

Communication

The findings of this exploratory analysis are principally communicated in the presentation associated with this document.