**Slide Presentation Script Outline**

Time limit = 5 mins, so roughly 650-750 words max.

<1>

Our project’s goal was to see if we could use topic modeling to predict stock performance by detecting evidence of management evasiveness.

A number of academic studies have used company conference call transcripts to extract useful signals that could predict future stock performance.

These calls generally have two sections, one that is totally scripted, and a Q&A where management answers questions. Some studies have concentrated on the Q&A section as being potentially more revealing.

Our hypothesis is that a potentially important signal to look for is the extent to which management appear to answer the questions posed vs avoid answering them, because this may be an indication they are not confident about something in the company’s prospects.

<2>

But how can you evaluate whether an answer relates to a question?

We chose the NLP technique of topic modeling which seeks to learn topics that represent the themes of documents. Our idea was to see if the topic of the question and the topic of the answer were similar.

There are a number of topic modeling techniques. We chose Latent Dirichlet Allocation, first proposed in 2003 and subsequently extended in a number of ways as it is a generative probabilistic model that learns topic distributions for documents, is well studied and has been successfully applied to short documents like our problem has.

<3>

We had to do extensive preprocessing to take conference call transcript files and parse them into question and answer pairs.

We stemmed our tokens

We found that we needed to apply stopwords lists to remove frequent words that would

We implemented the LDA model using the popular Python gensim package. An LDA model needs to have a number of topics specified, so we implemented searches across that.

We also found that some of the q/a pairs were too short for us to really expect them to have meaningful topics (though show good morning topics as example)

We developed initial topic models and evaluated them, then made changes to preprocessing and

We chose to use a measure called ‘u\_mass’ coherence to evaluate the quality of our learned topics.

Once we had quality topics we could evaluate them for our task.

<4>

We present results against our three goals

First, could we learn quality topics?

We believe we could.

There are three other

Are topics coherent, are documents associated with a relatively small number of topics, and are topics associated with relatively few words?

Here is an example of a learned topic with a high coherence score and the question that is most similar to that.

As a contrast, here is a topic with a low coherence score

Here’s a distribution of the number of topics that documents are associated with. In this model, most are associated with up to 9.

And here’s a distribution of the number of words that each topic is associated with. Most of the probability mass is concentrated in 25 words or less in each topic.

<5>

So if we were able to learn quality topics, are the similarities between question and answer of any use? A simple check on that is to look at the distribution of similarities vs those we would expect by chance, given these topics if the question and answer topic distributions are paired up at random.

This makes us also feel that topic similarity is a meaningful measure.

<6>

Finally, did topic similarity between question and answer help predict stock performance? Here we present results for our leading model vs the baseline. Unfortunately we do not see any statistically relevant relationship for the topic similarities. The baseline does not actually perform very well either.

<7>

So in conclusion we believe that we have shown topic modeling can be applied to earnings call transcripts, that meaningful topics can be extracted, that there is an interesting relationship between the topics of questions and answers.

We would see two main avenues for building on this work

First, expand the universe of stocks. We concentrated on the S&P 100 for a seven year period. Scaling the same techniques to the Wilshire 5000 would help in both learning more robust topics and getting coverage of smaller stocks which may have less efficient market coverage and hence more potential for surprise.

Second we could move the basis from word tokens to word embeddings. We chose to use tokens in our project as the corpus was challenging itself, but now we believe that we have demonstrated meaningful topics are present, it would be possible to try LDA with vectors instead.

Thanks for listening and we’re happy to take any questions.