Arun Varghese
Language and Information
Topic Assignment

1) **Resource presentation**
I would like to present a tutorial of Stan (http://mc-stan.org/), a programming language used for Bayesian statistial inference—a major component of some NLP techniques.

2) **Term paper/project**
A recent, influential paper in economics (Gentzkow and Shaprio 2010) examined political bias in the news. That it, they wanted to know whether newspapers delivered the news with inherent political slant or whether newspapers were instead catering to the demand of its customers for political bias. As an intermediate step, the authors constructed a method of political slant using NLP techniques. They compared the similarity of language used by various newspapers against the language used by Republicans and Democrats in the congressional record. Certain phrases tended to be more predominantly by Republicans or by Democrats. For example, Republicans are more likely to say, "death tax" while Democrats are more likely to say, "estate tax." If a newspaper was more likely to contain the phrase "death tax" than "estate tax," it scored as more Republican on that dimension.

I would like to go through a similar exercise with newspapers and political speech from Kerala, a state in South India. A majority of this discourse is in Malayalam, the local language. I understand Malayalam but I do not know very much about the politics of Kerala so it will be interesting to see what I can learn generally about Kerala politics through NLP methods.

Many Malayalam language newspapers are available online and I was also able to find some congressional-record type documents online. I however, do not know how difficult it is to process the Malayalam language. I was able to find some papers doing work with Malayalam but generally the literature did not seem very deep. The language is agglutinative which I know presents some unique problems. From a practical standpoint, I am also unsure how to work with non-Roman script.