Northeastern University

INFO 6105
Data Sci Eng Mth & Tools
Lecture 2 tm addendum for non-English text

16 January 2018

tm

- R text mining library
- `install.packages('tm")`
- `library(tm)`

Load hindi text & remove punctuation

- Assume `hindi.txt` contains Unicode for poem in hindi
  - Contained in RStudio Session folder
- `h <- Corpus(VectorSource(readLines("hindi.txt", n=1, encoding="UTF-8")))`

```
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] साजन!होलीआईहै!,सुखसेहँसना,जीभरगाना,मस्तीसेमनकोबहलाना,पर्वहोगयाआज-,साजन!होलीआईहै!,हँसानेहमकोआईहै!
```

```
> h <- tm_map(h, removePunctuation)
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] साजनहोलीआई हैसुखसेहँसनाजीभरगानामस्तीसेमनकोबहलानापर्वहोगयाआजसाजनहोलीआई हैहँसानेहम कोआई है
```

Removing stopwords from hindi frame

```
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] साजनहोलीआई हैसुखसेहँसनाजीभरगानामस्तीसेमनकोबहलानापर्वहोगयाआजसाजनहोलीआई  हैहँसानेहम
कोआई  है
```

```
> h <- tm_map(h, removeWords, c("साजनहोलीआई", "email"))
> inspect(h)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1]  हैसुखसेहँसनाजीभरगानामस्तीसेमनकोबहलानापर्वहोगयाआजसाजनहोलीआई  हैहँसानेहमकोआई  है
>
```