INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY

HYDERABAD

International Institute of Technology - Hyderabad
SMAI - SPRING 2019 Course Project

Project ID - 11

PREDICT MOVIE TAGLINE FROM TEXT SYNOPSIS

Team 31- PGSSP ML TECHNIES
Project guide : Satyam mittal
Project members : Jyothsna – Madan – Surekha - Karthikeyan

ABSTRACT

❑ To predict the tagline of a movie from given text synopsis.

❑ NLP Problem

❑ Abstractive Summarization

❑ Our Approach

GIVEN PROBLEM STATEMENT

- Predicting Movie Tagline for any Given Movie Synopsis

- Movie Tagline are Nothing but Abstract Summary of the movie.

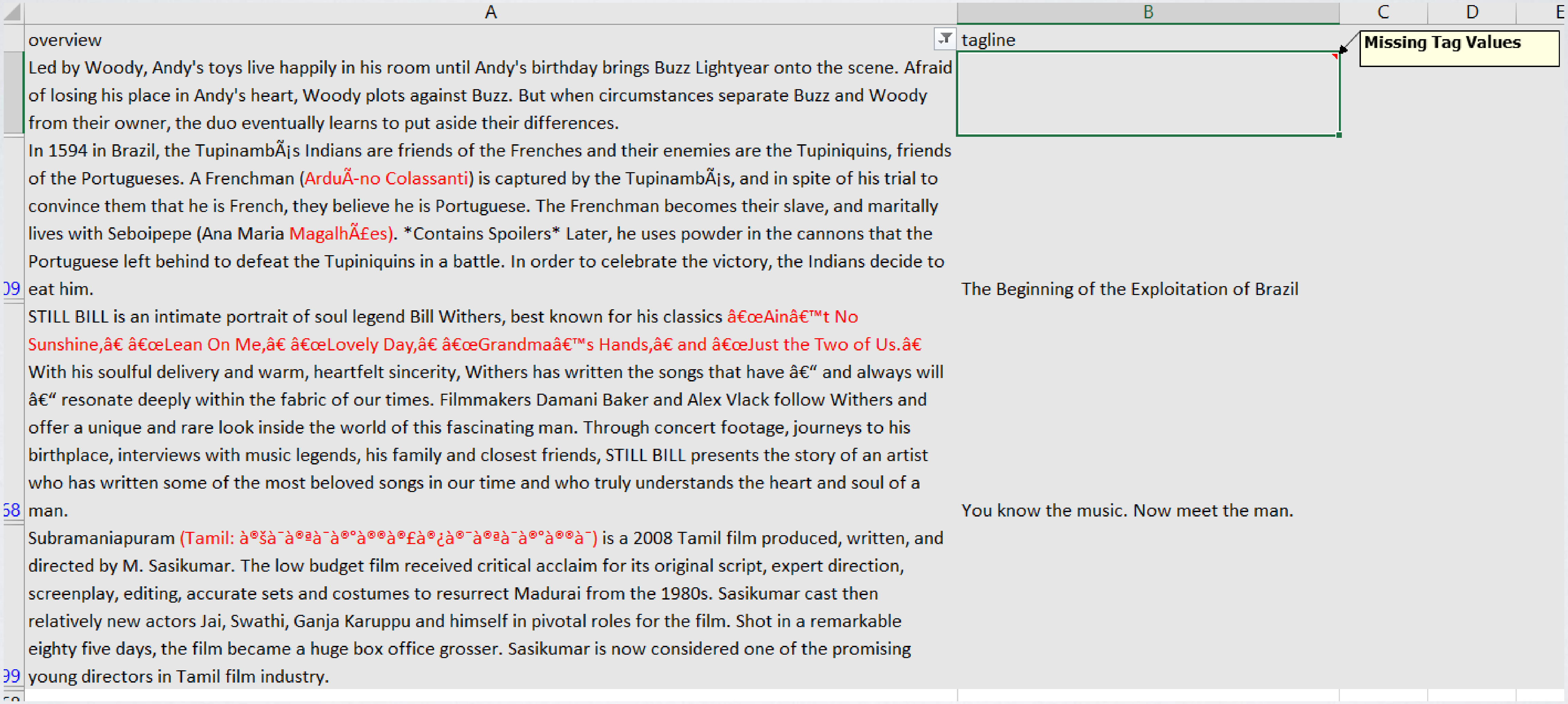
- Hence this is a Text Abstraction - NLP Problem

CHALLENGES

Data:

- Missing values (Either in Plot Synopsis or Movie tagline)
- Non UTF8 Text contents (such as tm or copyright symbols etc.)
- Non English Text Contents (Text written in regional Language fonts which are not a English char)

CHALLENGES



	A	B
	overview	tagline
	Led by Woody, Andy's toys live happily in his room until Andy's birthday brings Buzz Lightyear onto the scene. Afraid of losing his place in Andy's heart, Woody plots against Buzz. But when circumstances separate Buzz and Woody from their owner, the duo eventually learns to put aside their differences.	
)9	In 1594 in Brazil, the TupinambÃ¡s Indians are friends of the Frenches and their enemies are the Tupiniquins, friends of the Portugueses. A Frenchman (ArduÃ-no Colassanti) is captured by the TupinambÃ¡s, and in spite of his trial to convince them that he is French, they believe he is Portuguese. The Frenchman becomes their slave, and maritally lives with Seboipepe (Ana Maria MagalhÃ£es). *Contains Spoilers* Later, he uses powder in the cannons that the Portuguese left behind to defeat the Tupiniquins in a battle. In order to celebrate the victory, the Indians decide to eat him.	The Beginning of the Exploitation of Brazil
58	STILL BILL is an intimate portrait of soul legend Bill Withers, best known for his classics â€œAinâ€™t No Sunshine,â€ â€œLean On Me,â€ â€œLovely Day,â€ â€œGrandmaâ€™s Hands,â€ and â€œJust the Two of Us.â€ With his soulful delivery and warm, heartfelt sincerity, Withers has written the songs that have â€“ and always will â€“ resonate deeply within the fabric of our times. Filmmakers Damani Baker and Alex Vlack follow Withers and offer a unique and rare look inside the world of this fascinating man. Through concert footage, journeys to his birthplace, interviews with music legends, his family and closest friends, STILL BILL presents the story of an artist who has written some of the most beloved songs in our time and who truly understands the heart and soul of a man.	You know the music. Now meet the man.
)9	Subramaniapuram (Tamil: à®šà¯à®ªà¯à®°à®®à®£à®¿à®¯à®ªà¯à®°à®®à¯) is a 2008 Tamil film produced, written, and directed by M. Sasikumar. The low budget film received critical acclaim for its original script, expert direction, screenplay, editing, accurate sets and costumes to resurrect Madurai from the 1980s. Sasikumar cast then relatively new actors Jai, Swathi, Ganja Karuppu and himself in pivotal roles for the film. Shot in a remarkable eighty five days, the film became a huge box office grosser. Sasikumar is now considered one of the promising young directors in Tamil film industry.	

Missing Tag Values

CHALLENGES – CONT.

Training Model:

- Huge Volume of data to be processed
- Computation complexity leads to Long hours of training
- Risk of Losing training data while executing training for long hours – Alternatively we used Checkpoints to save progress of the training.

CHALLENGES – CONT.

Unknown words in Data set:

- There is always possibility of getting a Unknown word which is not part of word embedding matrix which we created. i.e Out of Vocabulary Words

- Examples : Name of Characters in the movie like Sivagami, Baahubali etc.

CHALLENGES – CONT.

Training Model:

- Huge Volume of data to be processed
- Computation complexity leads to Long hours of training
- Risk of Losing training data while executing training for long hours – Alternatively we used Checkpoints to save progress of the training.

IMPLEMENTATION DETAILS

- Bi Directional – Encoder decoder LSTM
- TensorFlow based implementation.
- Batch processing of Data

IMPLEMENTATION DETAILS , CONT.

- Word Vector Embedding .
  - Numberbatch word embedding* is built on:
    - ConceptNet 5.5,
    - GloVe,
    - word2vec,
    - Parallel text from OpenSubtitles 2016
    - * Refer GitHub Repository mentioned in the reference section
-

IMPLEMENTATION DETAILS , CONT.

- Intermediate Training Data/Weights is saved as Check points.
  - Can be Loaded back to system later.
  - More training data can be trained on top of existing Checkpoint.
  - Saved Check point can be distributed as Binary file to others to deploy the model

- Training Stop Criteria – Add Some explanation to this

DEMO

FUTURE WORKS

- To Extend this work to Indian Regional Languages
- Challenge with Language Corpus

REFERENCES – PART 1

- Predicting Movie Genres Based on Plot Summaries by Quan Hoang : https://arxiv.org/pdf/1801.04813.pdf

- Folksonomication: Predicting Tags for Movies from Plot Synopses UsingEmotion Flow Encoded Neural Network by Sudipta Kar, Suraj Maharjan and Thamar Solorio - https://aclweb.org/anthology/C18-1244

- Patent Abstract Summarization using Recurrent Neural Networks by Abhishek Jindal,Chirag Choudhary and Nile Hanov – https://github.com/ajindal1/Text_Summarizer_On_Patents/blob/master/project_report/Text_Summarization_project_NLP.pdf

REFERENCES – PART 2

- Data Source:
  - ❑ Found in Kaggle - https://www.kaggle.com/rounakbanik/the-movies-dataset
  - ❑ Filename : movies_metadata.csv
- Concept Net Number batch – https://github.com/commonsense/conceptnet-numberbatch
- Glove - https://nlp.stanford.edu/projects/glove/

REFERENCES – PART 3

- Rouge Score: https://stats.stackexchange.com/questions/301626/interpreting-rouge-scores
- Performance metrics - https://nlpprogress.com/english/summarization.html
- RNN – Sequence to Sequence Model : https://towardsdatascience.com/seq2seq-model-in-tensorflow-ec0c557e560f

Thank you!