Phase 1 Report

Group 3

Project Title

- Hash-tag Generation for Social Media Content

Mentor

- Shashank Gupta

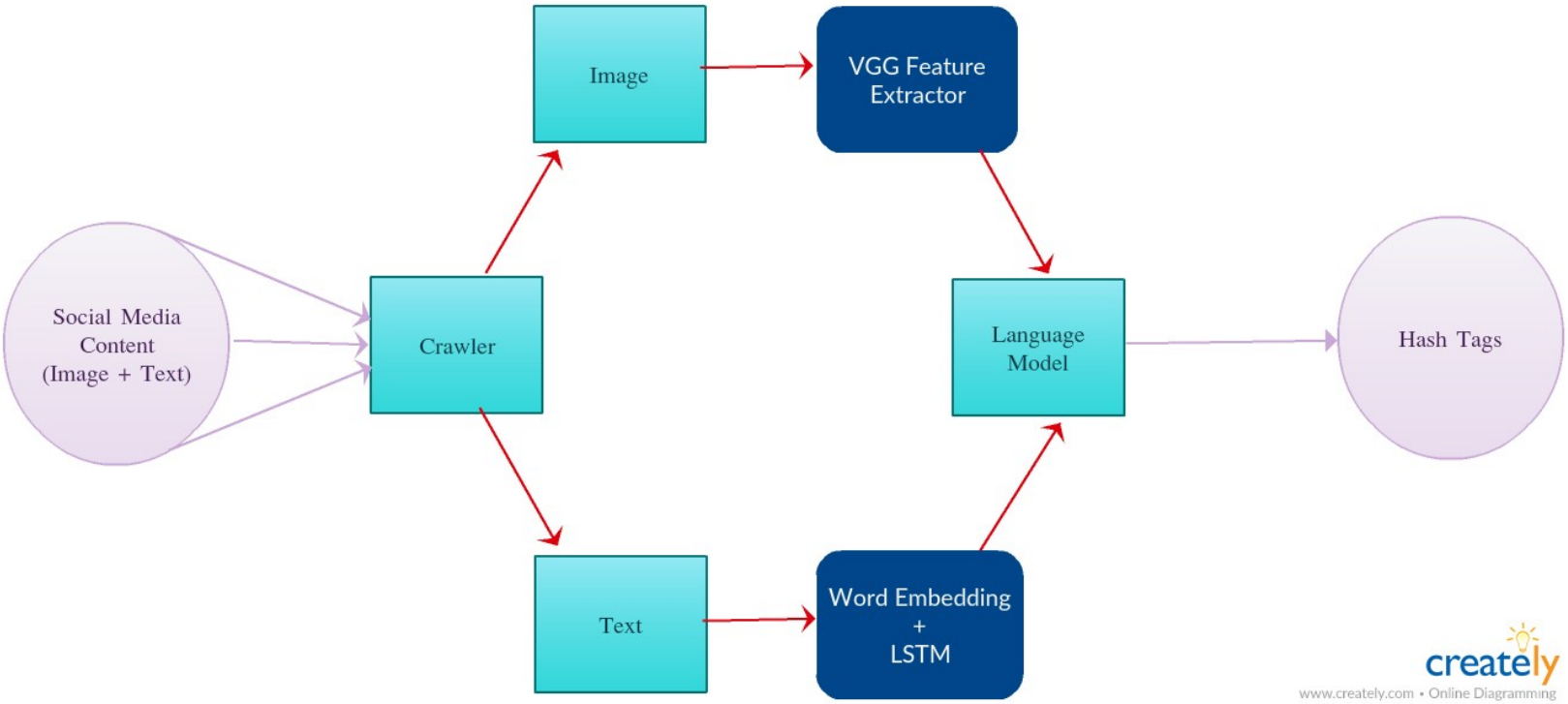
Members

- Harshil Jain
- Syed Ahmad
- Kaleemullah Mohammed
- Krishna Chaitanya Pappu

Project Description

The project takes Social media content having text and images and generates hash-tags describing the content. It is not a classification problem where we assign one of the predefined set of labels, rather we generate new hash-tags each time using a trained language model.

System Design



Phase 1

Crawling Data

We are crawling various tweets having text as well as images in them from various user accounts. It is an infinite crawling process, abiding by the request limits of tweepy API. The statistics as of today.

Number of Tweets	Number of Images	Number of Hash tags
15 lakhs	15 lakhs	~18 lakhs (non-distinct)

Data Model

The tweets are stored in a csv file, where the format is

<Id>,<Time stamp>,<Text>,<URL to the image>

The images are processed post downloading them using the URLs.

The text section is processed to separate plain text and hash tags. And then , the plain text is processed to remove non-alpha numeric characters and then for each tweet, we are making two dictionaries

word_to_id - It is normal index {word:id}

id_to_embed - It stores, the id of a word and its corresponding Glove embedding of 100 dimensions {id:embedding}.

The Glove embedding for twitter were obtained from internet.

Model

Text Feature Extraction

The processed text is sent in as sequence of length 15 where we get the word ids and store them in a matrix. If the tweet has more than 15 words only the first 15 are considered and if the tweet has less than 15 words then it is padded with <unk> which is also the case with words not found in the glove embeddings. Then it is passed to an LSTM model, which is similar to a word level language model and the model outputs a 1000 dimensional vector, which will be fed into the character level language model for hashtag generation along with the Image features.

Phase 2 Objectives

Crawling Data

The crawling goes on the in the background till we hit required number of tweets. This is not a major component of this phase.

Model

Completing the character level language model and feeding it inputs from both the media and training it.

Once, the model is up and ready tweaking its parameters and layers to optimize the results.

Challenges

- Each tweet has one or more than one hash tags, how to give labels and prepare the training set is a challenge.
- We have come up with 2-3 approaches, trying them on a sample data and finalizing one of them for the final training of the model.