**Script Compression**

**Word importance**

A word in a script is related to the sentence it is in as well as the scene.

Importance of a word is defined by:

1. Inherent importance:

* POS tag (Verb, Adjective, Noun, etc)
* Word Frequency

2. Relative importance:

* importance of word in the sentence
* relative importance of the sentence wrt other sentences in the paragraph
* if the sentence is a part of AC line or dialogue what is the relative importance of the sentence in the AC line or dialogue.
* Relative importance of the AC line or dialogue in the scene.
* followed by the relative importance of the scene in the film

Word importance (word) = Inherent importance + Relative importance

Stop words: The importance of stop words will depend on the sentence itself. If the important words are getting removed from the sentence then stop words has to be removed.

**Removal impact**

Removal impact of words will depend on:

* The sentence from which it has been removed and the sentences followed by it.
* Removing a Verb or Noun will impact the whole sentence.
* If the word is a speaker – the dialogue will be impacted along with the dialogues followed by it. So, the initial dialogues will have higher impact.
* If the word is a slug line, whole scene will be impacted.
* Removing a scene will impact the scenes followed by it.

Note: When calculating removal impact or importance of a scene, scenes with more no. of AC/dialogues will dominate. Thus an average value should be calculated i.e total removal impact or importance divided by the total no. of words in the scene.

**How to remove words**

* Words should be removed scene by scene
* If the words are removed in increasing order of removal impact, it may happen that words may start getting removed from the last scene and the counter stops till the initial scenes are near, thus resulting in word removal of only words in the end.
* For a particular scene a threshold value could be set for the words that’ll be removed.
* Words should be removed one scene at a time, e g. first from the first scene with less removal impact should be removed, followed by the next scenes.

Factors:

3. Word position – Initial words carry more info. Eg: Kush is driving to the gym. - here kush is driving has more importance.

6. Semantic similarity – using PMI

9. Stop words: Stop words don’t contribute much

But each sentence in a document has different importance for identifying the content of the document.

Thus, by assigning a different weight according to the importance of the sentence to each term, we can achieve better results.

Among text summarization techniques, there are statistical methods and linguistic methods.

The importance of each sentence is measured by term weights.

Measuring the importance of sentences:

First, the sentences, which are more similar to the title, have higher weights

In the next method, we first measure the importance of terms by TF, IDF, and chi-square statistics values.

Then we assign the higher importance to the sentence with more important terms.

we first measure the importance values of terms by TF, IDF, and v2 statistics values, and then the sum of the importance values of terms in each sentence is assigned to the importance value of the sentence. In this method, the importance value of a sentence Si in a document d is calculated.

1. Term frequency = No. of times a term occurs in a document.

TF(t,d) = count of t in d/ no. of words in d

2. Document Frequency: The importance of document in whole set of documents. TF is frequency counter for a term t in document d, whereas DF is the no. of documents in which the word is present.

DF(t) = occurrence of t in documents

3. IDF = There are lots of words like ‘is’, ‘of’, etc that appear a lot of times but have a little importance. Thus IDF diminishes the weights of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

IDF = total no of documents/ no of documents in which term ‘t’ occurs

string similarity – string is a sequence not a sentence.

**Subject:** typically a noun or pronoun—the person, place or thing  
**Verb:** the action or state of being  
**Object:** the word or group of words influenced by the verb