**KEYWORD/KEYPHRASE IDENTIFICATION**

**(or TEXT SEGMENTATION) OF THE SEARCH QUERY**

By: **Arpit Singh**

**Introduction:**

When we write a search query which is a short sentence, we want to know what the keywords/keyphrases are in that sentence which will then be used to get the search results for the corrresponding query. For getting better search results, we need to segment the text in the most meaningful way (for the best identification of keywords/keyphrases) in the given search query.

For example, if the search query is 'New Delhi in India', then its text chunks are 'new delhi' and 'india' and not 'new', 'delhi' and 'india'.

During my project, I worked on finding the best segmentation of the text in the search query.

However, in this, the traditional approaches (like TF\*IDF, etc.) do not work properly as a short sentence does not have enough statistics. Also, the search query may not properly follow the syntax of the written language.

So I used semantic relation and cooccurence frequency between the phrases (and also POS tags) as parameters to obtain the best segmentation of the search query. The detailed algorithm is listed below.

**Thoughts behind the algorithm:**

If the phrases are highly semantically related to each other, then there is a higher chance for those phrases to exist separately and not as a single keyphrase. So we try to find that combination of keyphrases such that the **semantic relation** between those keyphrases overall is maximum.

However, we also need to account for the **cooccurence frequency** between those phrases because the algorithm used by the word2vec library (which is used for getting the semantic relation) uses CBOW algorithm and thus sometimes higher cooccurence frequency leads to higher semantic relation which results in the phrases to get separated in my algorithm when instead they should occur together (as they have higher cooccurence frequency and thus higher probability to occur together). So the final relation is decided by a combination of them in which the semantic relation is the major parameter and cooccurence frequency is used to increase or decrease its impact depending on how much less or more the phrases cooccur in the training data. The POS tags are also considered for the case when there are two words.

**Algorithm:**

We use the corpus given by the expert and also the 'text8' corpus (that comes along with the word2vec library) and train the data using the **word2vec** library. This library is used to get semantic relation and coccurrence between any two phrases.

Firstly, the expert data is merged into a single file and the data is cleaned. Then the stopwords are removed. I wrote a script that calculates the distance between the two phrases which helps us give the semantic relation between the words and also caculates the frequency of the cooccurnce of the combination of the two phrases which helps me to get the cooccurence frequency of the two phrases.

Then its possible contiguous subsequence phrases are found. Then we store the semantic relation, cooccurence relation and final relation between these obtained phrases in a **graph**.

After this, the maximum weighted edge is found and then it removes all nodes that are disconnected with the picked nodes corresponding to the max weighted edge (this is done to remove words like, for example, 'natural' to come again if 'natural resources' has already been selected as the keyword). At the same time, it removes all edges that are linked to the deleted nodes. This process is repeated until no edges can be selected. The nodes that are finally picked are the keyphrases of the search query.

**Special case of two words**: In this case, my algorithm always prints them as separate words (in the algorithm, the nodes of the max weighted edge are taken, so there will be atleast 2 phrases) and also the semantic relation is not a reliable parameter. So the combination of **POS tags** and cooccurence frequency is used to make the algorithm work properly for this case. For example, if they have a higher coccurence frequency or their POS tags are of 'adjective noun' or 'noun noun', they are more likely to occur together.

To understand the algorithm in a better manner, please have a look at my code.

**Results:**

In this, there are two things that should be kept in mind.

1. Evaluating the text segmentation is slightly dependent on how a person thinks they should be segmented (for example, 'natural resources in india' should be chunked as 'natural resources' and 'india' by some person but another may argue that it should be chunked as 'natural', 'resources' and 'india' because the word 'natural' can be asociated witha lot of other things other than the word 'resources').

2. The results are heavily dependent on the data (for example, if we take a data in which 'united states' does not appear and rather phrases like 'united india' and 'united team' occur much, then this will provide united and states as different chunks and not as a single chunk).

I ran the code for the following queries and their results are as follows:

new delhi india

new delhi ( JJ NN )

india ( NN )

appalachicola river in florida

appalachicola ( NN )

river ( NN )

florida ( NN )

rainwater harvesting

rainwater harvesting ( NN VBG )

bee wax and royal jelly

bee ( NN )

wax ( NN )

royal jelly ( NN RB )

bija yatra

bija yatra ( NN NN )

iron ores of the kallakurchi

iron ( NN )

ores ( NNS )

kallakurchi ( NN )

self reliance in food

self reliance ( NN NN )

food ( NN )

Gene Campaign

gene campaign ( NN NN )

prices of chemical pesticide and fertilizers

prices ( NNS )

chemical ( NN )

pesticide ( NN )

fertilizers ( NNS )

valedictory address by Prasant Mohanty

valedictory ( NN )

address ( NN )

prasant mohanty ( JJ NN )

budget allocation for revival of agriculture

budget ( NN )

allocation ( NN )

revival ( NN )

agriculture ( NN )

Green Revolution on food security

green revolution ( JJ NN )

food security ( NN NN )

indigenous variety of rice by farmers

indigenous ( JJ )

variety ( NN )

rice ( NN )

farmers ( NNS )

Mukurunda tribals of Rajasthan

mukurunda ( NN )

tribals ( NNS )

rajasthan ( NN )

manufacture of cortisone and sex hormones

manufacture ( NN )

cortisone ( NN )

sex hormones ( NN NNS )

morphological characteristics of S. xanthocarpum

morphological ( JJ )

characteristics ( NNS )

s xanthocarpum ( NN NN )

strength of experimental ointment

strength ( NN )

experimental ( JJ )

ointment ( NN )

evaluation of woundhealing activity

evaluation ( NN )

woundhealing ( VBG )

activity ( NN )

granular tissue of the control group of animals

granular ( JJ )

tissue ( NN )

control ( NN )

group ( NN )

animals ( NNS )

Methanol extract

methanol extract ( NN NN )

growth of dairy industry in india

growth ( NN )

dairy ( NN )

industry ( NN )

india ( NN )

murrah buffalo

murrah buffalo ( NN NN )

livestock fairs in Rajasthan

livestock ( NN )

fairs ( NNS )

rajasthan ( NN )

malpractices observed in marketing of Cattle

malpractices ( NNS )

observed ( VBD )

marketing ( NN )

cattle ( NNS )

Poor milk yielding cows

poor ( JJ )

milk ( NN )

yielding ( NN )

cows ( NNS )

Assembling and distribution of animals

assembling ( VBG )

distribution ( NN )

animals ( NNS )

shelter facility for animals

shelter ( NN )

facility ( NN )

animals ( NNS )

check the spread of these contagious diseases

check ( VB )

spread ( NN )

contagious ( JJ )

diseases ( NNS )

support of ford foundation

support ( NN )

ford foundation ( NN NN )

development of livestock sector

development ( NN )

livestock ( NN )

sector ( NN )

MARKETS FOR ORGANIC PRODUCE IN JHARKHAND

markets ( NNS )

organic ( JJ )

produce ( NN )

jharkhand ( NN )

administrative reforms in livestock fairs

administrative ( JJ )

reforms ( NNS )

livestock fairs ( NN NNS )

science of yoga

science ( NN )

yoga ( NN )

Cessation of modification of Chitta

cessation ( NN )

modification ( NN )

chitta ( NN )

Yoga and naturopathy

yoga ( NN )

naturopathy ( JJ )

sustainable farming practices

sustainable ( JJ )

farming ( NN )

practices ( NNS )

reliance on animal draught power

reliance ( NN )

animal ( NN )

draught power ( NN NN )

Facilitating land preparation

facilitating ( VBG )

land ( NN )

preparation ( NN )

situation of conservation agriculture in Zambia

situation ( NN )

conservation agriculture ( NN NN )

zambia ( NN )

environmental degradation in zambia

environmental ( JJ )

degradation ( NN )

zambia ( NN )

seedling selection from Chakaiya

seedling selection ( VBG NN )

chakaiya ( NN )

Barbados Cherry in Kerala

barbados cherry ( NN NN )

kerala ( NN )

aromatic rice cultivars

aromatic rice ( JJ NN )

cultivars ( NNS )

X-rays diffraction

x rays ( JJ NNS )

diffraction ( NN )

Environmentalism as a faith system

environmentalism ( NN )

faith ( NN )

system ( NN )

uniqueness of Sangham

uniqueness ( NN )

sangham ( NN )

renaissance of medicinal plants

renaissance ( NN )

medicinal ( JJ )

plants ( NNS )

Dry powder of suran

dry ( JJ )

powder ( NN )

suran ( NN )

genetic and species diversity

genetic ( JJ )

species ( NNS )

diversity ( NN )

decoction of arush with sugar

decoction ( NN )

arush ( NN )

sugar ( NN )

Fenugreek seeds and lukewarm water

fenugreek seeds ( NN NNS )

lukewarm water ( JJ NN )

Paste of leaves of neem

paste ( NN )

leaves ( NNS )

neem ( NN )

treatment of normal attack of paralysis

treatment ( NN )

normal ( JJ )

attack ( NN )

paralysis ( NN )

Cure of moles, blain and blister

cure ( NN )

moles ( NNS )

blain ( VBP )

blister ( NN )

indigenous practices for healthcare

indigenous ( JJ )

practices ( NNS )

healthcare ( NN )

management of seed bank

management ( NN )

seed bank ( NN NN )

Community Rights

community ( NN )

rights ( NNS )

Village Council of Devrampalli in Medak

village ( NN )

council ( NN )

devrampalli ( NN )

medak ( NN )

seeds of tender fruits

seeds ( NNS )

tender ( NN )

fruits ( NNS )

recommended requirement of vegetables

recommended ( JJ )

requirement ( NN )

vegetables ( NNS )

agents in pharmaceutical preparations

agents ( NNS )

pharmaceutical ( JJ )

preparations ( NNS )

flavouring plants of Assam

flavouring ( VBG )

plants ( NNS )

assam ( NN )

wild flora in the Deliblato Sand

wild ( JJ )

flora ( NNS )

deliblato sand ( NN NN )

rural populations in Serbia

rural ( JJ )

populations ( NNS )

serbia ( NN )

Botanical diversity of plants

botanical ( JJ )

diversity ( NN )

plants ( NNS )

tuberous species of Kolli Hills

tuberous ( JJ )

species ( NNS )

kolli hills ( NN NNS )

staple food of Western Ghats

staple ( JJ )

food ( NN )

western ( JJ )

ghats ( NNS )

tribal sects of India

tribal ( JJ )

sects ( NNS )

india ( NN )

Mining Bees

mining ( NN )

bees ( NNS )

Toxicity of Pesticides to Bees

toxicity ( NN )

pesticides ( NNS )

bees ( NNS )

bee pollination

bee pollination ( NN NN )

weeds of kanyakumari

weeds ( NNS )

kanyakumari ( NN )

enumeration of medicinally important weeds

enumeration ( NN )

medicinally important ( RB JJ )

weeds ( NNS )

application of poultices

application ( NN )

poultices ( NNS )

leaf decoction

leaf decoction ( NN NN )

WEAVING TECHNIQUE OF DOOR SCREEN

weaving ( VBG )

technique ( NN )

door ( NN )

screen ( NN )

network perspective of entrepreneurship

network ( NN )

perspective ( NN )

entrepreneurship ( NN )

Social networks of master weavers

social ( JJ )

networks ( NNS )

master weavers ( NN NNS )

Hypotheses for structural embeddedness

hypotheses ( NNS )

structural embeddedness ( JJ NN )

Government interventions in handloom

government ( NN )

interventions ( NNS )

handloom ( NN )

Human Capital on performance

human capital ( JJ NN )

performance ( NN )

the analysis and the findings of Long

analysis ( NN )

findings ( NNS )

long ( JJ )

weaving door screen

weaving ( VBG )

door ( NN )

screen ( NN )

weaving of saree, lungi and napkin

weaving ( NN )

saree ( JJ )

lungi ( NN )

napkin ( NN )

making of Parda

making ( NN )

parda ( NN )

village population of Dhalapathar

village ( NN )

population ( NN )

dhalapathar ( NN )

Wild and Cultivated Species of Cotton

wild ( JJ )

cultivated ( JJ )

species ( NNS )

cotton ( NN )

Characters of breeding value

characters ( NNS )

breeding ( VBG )

value ( NN )

Advances in Applied Science

advances ( NNS )

applied ( JJ )

science ( NN )

genomic constitution of banana

genomic ( JJ )

constitution ( NN )

banana ( NN )

wheat production in India

wheat ( NN )

production ( NN )

india ( NN )

wheat landraces of Oman

wheat ( NN )

landraces ( NNS )

oman ( NN )

livelihood projects in orissa

livelihood ( NN )

projects ( NNS )

orissa ( NN )

Weed management through salt application

weed ( NN )

management ( NN )

salt ( NN )

application ( NN )

invasion of weeds

invasion ( NN )

weeds ( NNS )

traditional practices to use salt

traditional ( JJ )

practices ( NNS )

use ( VB )

salt ( NN )

vetiver system

vetiver system ( NN NN )

Microbial analysis of manure

microbial ( JJ )

analysis ( NN )

manure ( NN )

application of vermiwash

application ( NN )

vermiwash ( NN )

use of panchgavya in cauliflower

use ( NN )

panchgavya ( NN )

cauliflower ( NN )

organic farm at palampur

organic ( JJ )

farm ( NN )

palampur ( NN )

compost tea

compost tea ( NN NN )

hazards of vegetable dyes

hazards ( NNS )

vegetable ( JJ )

dyes ( NNS )

annual production of dye in Jaipur

annual ( JJ )

production ( NN )

dye ( NN )

jaipur ( NN )

composition of natural dye

composition ( NN )

natural ( JJ )

dye ( NN )

art of making vegetable dyes

art ( NN )

making ( VBG )

vegetable ( JJ )

dyes ( NNS )

review of sri

review ( NN )

sri ( NN )

Jessour technique in Tunisia

jessour ( NN )

technique ( NN )

tunisia ( NN )

breeding of the Vechur

breeding ( NN )

vechur ( NN )

Utilization and Technology of Water in Agriculture

utilization ( NN )

technology ( NN )

water ( NN )

agriculture ( NN )

Sustainable Agriculture

sustainable agriculture ( JJ NN )

Application of cow dung

application ( NN )

cow ( NN )

dung ( NN )

water management in farmland

water ( NN )

management ( NN )

farmland ( NN )

rise in groundwater level

rise ( NN )

groundwater ( NN )

level ( NN )

Deterioration of pasture land

deterioration ( NN )

pasture ( NN )

land ( NN )

flood control and utilization of water

flood control ( NN NN )

utilization ( NN )

water ( NN )

mangrove forests in Odidha

mangrove ( NN )

forests ( NNS )

odidha ( NN )

pest management in Assam

pest ( JJS )

management ( NN )

assam ( NN )

produtive wetlands

produtive ( JJ )

wetlands ( NN )

use of Aegiceras corniculatum

use ( NN )

aegiceras ( NNS )

corniculatum ( NN )

phytomedicinal knowledge of Bhotias of Dharchula

phytomedicinal ( JJ )

knowledge ( NN )

bhotias ( NN )

dharchula ( NN )

cosmos flowers

cosmos ( NN )

flowers ( NNS )

Preparation of yarn for dyeing

preparation ( NN )

yarn ( NN )

dyeing ( VBG )

Evaluation of colour fastness

evaluation ( NN )

colour ( NN )

fastness ( NN )

Traditional Phytotherapy among Karens

traditional ( JJ )

phytotherapy ( NN )

karens ( NNS )

Control of Gundhi bug

control ( NN )

gundhi bug ( NN NN )

threat status of endemic grasses

threat ( NN )

status ( NN )

endemic ( JJ )

grasses ( NNS )

production trend of coarse cereals

production ( NN )

trend ( NN )

coarse cereals ( NN NNS )

MARKETING PROBLEMS OF MICRO ARTISAN ENTERPRISES

marketing ( NN )

problems ( NNS )

micro artisan enterprises ( JJ JJ NNS )

taxonomic studies of Poaceae

taxonomic ( JJ )

studies ( NNS )

poaceae ( NN )

Dimeria, Eulalia and Themeda

dimeria ( NNS )

eulalia ( NNS )

themeda ( NN )

fodder crops

fodder ( NN )

crops ( NNS )

Loom Material dealers

loom ( NN )

material ( NN )

dealers ( NNS )

Role of seriFed in silk industry

role ( NN )

serifed ( NN )

silk ( NN )

industry ( NN )

muga silk

muga silk ( NN NN )

silk production in Asia

silk ( NN )

production ( NN )

asia ( NN )

appropriate cholesterol levels in heart

appropriate ( JJ )

cholesterol ( NN )

levels ( NNS )

heart ( NN )

properties of pomegranate seeds

properties ( NNS )

pomegranate ( NN )

seeds ( NNS )

fibre extraction of Sunnhemp

fibre extraction ( JJ NN )

sunnhemp ( NN )

phytosociological analysis of a plant

phytosociological ( JJ )

analysis ( NN )

plant ( NN )

Natural remedies for heart diseases

natural ( JJ )

remedies ( NNS )

heart ( NN )

diseases ( NNS )

Packaging and marketing of kokum products

packaging ( NN )

marketing ( NN )

kokum ( NN )

products ( NNS )

Cultivation of cabbage

cultivation ( NN )

cabbage ( NN )

lignans in millets

lignans ( NNS )

millets ( NNS )

DISASTER MANAGEMENT OF LEPCHA COMMUNITY

disaster ( NN )

management ( NN )

lepcha ( NN )

community ( NN )

effect of zhuming

effect ( NN )

zhuming ( VBG )

**Analysis of the results:**

For the above queries all of them give proper results according to me except the following (out of the above 150 queries, 128 queries produced correct results) :

iron ores of the kallakurchi

Currently: iron, ore, kallakurchi

Should be: iron ore, kallakurchi

Mukurunda tribals of Rajasthan

Currently: Mukurunda, tribe, Rajasthan

Should be: Mukurunda tribe, Rajasthan

granular tissue of the control group of animals

Currently: granular, tissue, control, group, animals

Should be: granular tissue, control group, animals

livestock fairs in Rajasthan

Currently: livestock, fairs, Rajasthan

Should be: livestock fairs, Rajasthan

Poor milk yielding cows

Currently: poor, milk, yielding, cows

Should be: poor, milk yielding, cows

Community Rights

Currently: community, rights

Should be: community rights

Village Council of Devrampalli in Medak

Currently: village, council, devrampalli, medak

Should be: village council, devrampalli, medak

staple food of Western Ghats

Currently: staple, food, western, ghats

Should be: staple, food, western ghats

Mining Bees

Currently: mining, bees

Should be: mining bees

WEAVING TECHNIQUE OF DOOR SCREEN

Currently: weaving, technique, door, screen

Should be: weaving, technique, door screen

Government interventions in handloom

Currently: government, interventions, handloom

Should be: government interventions, handloom

weaving door screen

Currently: weaving, door, screen

Should be: weaving, door screen

Advances in Applied Science

Currently: advances, applied science

Should be: advances, applied, science

Weed management through salt application

Currently: weed, management, salt, application

Should be: weed management, salt application

hazards of vegetable dyes

Currently: hazards, vegetable, dyes

Should be: hazards, vegetable dyes

composition of natural dye

Currently: composition, natural, dye

Should be: composition, natural dye

art of making vegetable dyes

Currently: art, making, vegetable, dyes

Should be: art, making, vegetable dyes

Jessour technique in Tunisia

Currently: jessour, technique, Tunisia

Should be: jessour technique, tunisia

Application of cow dung

Currently: application, cow, dung

Should be: application, cow dung

use of Aegiceras corniculatum

Currently: use, aegiceras, corniculatum

Should be: use, aegiceras corniculatum

MARKETING PROBLEMS OF MICRO ARTISAN ENTERPRISES

Currently: marketing, problems, micro artisan enterprises

Should be: marketing, problems, micro artisan, enterprises

Loom Material dealers

Currently: loom, material, dealers

Should be: loom material, dealers

**Conclusion:**

The algorithm produces quite good results . But this is a kind of a problem in which there is still a huge potential to improve the results which can be done by taking more data and/or taking user feedback and/or identifying and considering more factors and their combinations.

**References:**

1. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/conceptualization.pdf>

2. <https://code.google.com/p/word2vec/>