GREAT STEP 2018
SAFETY DATA ANALYTICS

PROBLEM STATEMENT

Problem Statement

InsAnalyzer would like to analyze the queries that the consumers of an FMCG company type in through an online platform and respond accordingly with relevant information.

The input data has 3 components / attributes / columns:

1. Various keywords that are primary concern of the consumers or seeking for information. These can be put under various categories such as brand, product line and sub-product line, usage, etc;
2. Examples of inputs (searched phrases) that the consumers may type in
3. Examples of formation of composite keywords (phrases) out of keywords and thus enhancing the list of keywords (‘shoe polish’ out of ‘shoe’ and ‘polish’)
4. Examples of Synonyms

A)

You need to create an optimal set of categories (buckets) in the range of 10 to 20 (e.g. brand, product line, price, region, etc.) and an optimal set of sub-categories under each category (e.g. wellness, skin care, makeup, etc under ‘product’ category; Lux, Hamam, Liril, Lakme, Lipton, etc sub-categories under ‘brand’ category; cheap, expensive under ‘price’ category; etc.) out of the Keywords; examples of the final attributes /keywords under a sub-category such as ‘loofah’, ‘soap’, ‘shower gel’ etc under ‘Lux’ sub-category. And you need to put all the key words in one or many categories and/ or sub-categories. The keywords may belong to different categories or sub-categories (e.g. polish can be under makeup ‘nail polish’ and also under accessories ‘shoe polish’).

B)

Whenever a consumer types in a phrase for information, you need to estimate or compute the closeness (co-relationship) of that phrase with various categories (buckets) and sub-categories. The outcome (closeness to various categories/ sub-categories) should be produced in order of a closeness values. You are supposed to define a quantitative metric for closeness or association.

C)

Consumers of various demography (e.g. academic, ethnic origin, geography/ region, profession, economic status, etc) may type in that they are used to or use colloquially (e.g. people from USA use ‘kidding’ to mean same as Asians use ‘joking’ or usage of the words ‘cab’ and ‘taxi’ in different regions). Your algorithm should be robust enough to take care of this kind of similar words used for same meaning.

D)

More importantly, consumer may type in a phrase having one or more keyword(s) NOT available in the given / current database. You need to develop an algorithm to put or assign those new keyword(s) in various categories and / or sub- categories by identifying relevant SYNONYMS applicable in the said context. Post that you should repeat step (B) and produce the output in order of closeness to various categories and/ or sub-categories.

E)

The performance of your algorithm will be tested by our test dataset. Hence, your algorithm should be generic enough to pass our test data and achieve sufficiently high accuracy level**.**

Note 1: You may require applying more than one supervised and / or unsupervised learning techniques for the same.

Note 2: Substantial weightage will be given to the overall approach (solution design) to solve this business problem. The approach note and requisite flowchart to be provided for evaluation.

Note 3: The primary objective is to define a closeness or similarity or association metric of the phrases to various categories and sub categories so that any input phrase can be assigned to various categories and sub categories with estimated (computed) closeness or association values.