You are given a set of phrases (let us call them keywords), each of 1-4 words. Group the keywords into clusters (groups of keywords) by picking keywords which are similar to each other.

There are multiple ways to cluster(group) them
1) using single keyword as the cluster name. Eg cluster name "gifts" - this cluster should have all keywords which has gift in it, christmas - should have all keywords which have christmas in it and so on
2) considering 2 words as the cluster name: gift ideas, gifts, christmas, craft. All keywords which have the 2 word cluster name appearing in sequence (phrase matching) in the keyword fall into that cluster

It is possible that the same keyword is part of 1 or more clusters.

To pick the cluster name: You can pick the most popular 1 to 3 word sequence from the given keyword set

Ideal Scenario:
The Cluster allocation is considered to be better if the following criteria is met
1. We should have only the larger clusters [with a cut off size on keywords]
2. If gift and gifts are considered similar. A Stemmer can be used for the same
3. If stop words, synonyms can be provided to the algorithm then it would make better clusters
4. If we can establish a hierarchy in the clusters. eg. [gifts] as parent -> [christmas, birthday, dad] as children. [ideas] as parent -> [christmas, birthday, dad]

Tech considerations
1. Maintain object oriented code
2. Save the keywords, their attributes, the output - clusters in a data store [database]
3. Time and memory consumption of the script should be considered as few hundred thousand keywords might be given as input for clustering.

Sample Input & Output
Eg input set of keywords:
best dad gifts
best gifts
birthday gift ideas
birthday gifts
boss gift ideas
boyfriend gift ideas
christmas craft ideas
christmas for dad
christmas gifts for dad
christmas gift ideas
christmas gifts
christmas gifts dad
cool gifts
craft gift ideas
craft ideas
craft room ideas
crafts

Sample Clusters:
gifts: best dad gifts, best gifts, birthday gifts, christmas gifts for dad, christmas gifts, christmas gifts dad, cool gifts
gift ideas: birthday gift ideas, boss gift ideas, boyfriend gift ideas,christmas gift ideas, craft gift ideas
christmas: "christmas craft ideas", "christmas for dad", "christmas gifts for dad", "christmas gift ideas", christmas gifts, christmas gifts dad

and so on.

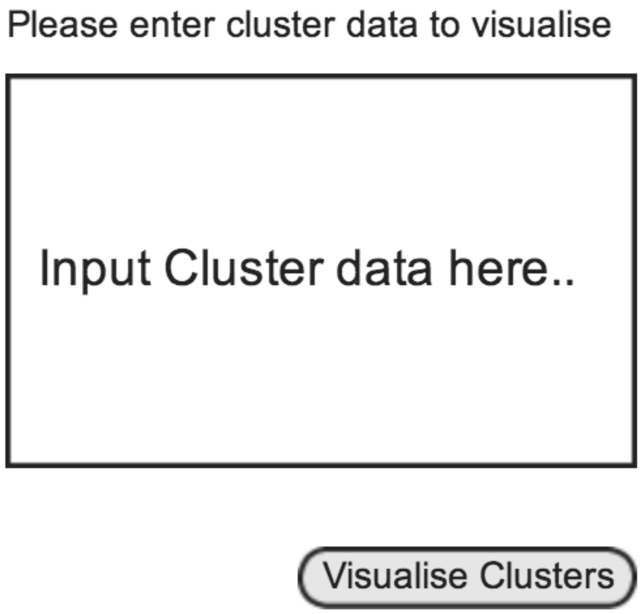
Backend Project
The program should read any data set from a file, find most popular common words as cluster names

and **print top N clusters.**
User can be prompted for the dataset filename & N or this can be read from a configuration file.

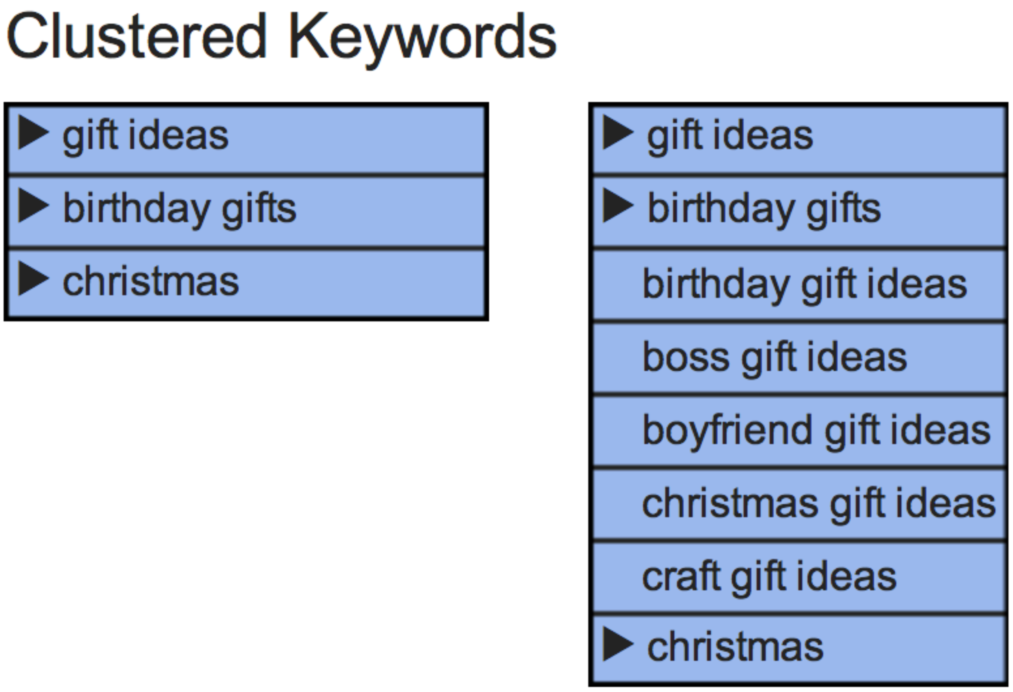
The data set is attached in another file.

UI Project

Come up with a way for the user to browse the clusters.
Input: grouped clusters along with keywords (sample i/p - in a text box let user enter the output of the backend program)


Please enter cluster data to visualise
Input Cluster data here..
Visualise Clusters

Output: A collapsable div of cluster names. When user clicks on the cluster name it should display the keywords.



Desired add on (optional): broad matching based on individual tokens. keyword "dad gifts" can match "best dad gifts", "christmas gifts for dad",
christmas gifts dad"
For the UI: you can integrate http://stacks.math.columbia.edu/tag/01WC/graph/collapsible