

**Goal:** The goal of this task is to validate the keywords assigned to a text.

**Introduction:** I am working on finding a Skill Model associated with a course. The data is taken from “Introduction to Biology” course which is available [here](#).

**Approach:**

I created clusters of text from the content in this course and then extracted keywords from each cluster. My hypothesis is that the keywords associated with each cluster represent the skills associated with them. I want to validate this hypothesis. I want to check whether keywords (skills) associated with the clusters are correct or not.

**Important Points:**

1. The clusters are available [here](#).
2. The clusters have been stored as set of 16 files. So, there are 16 clusters in total
3. The file names are “cluster0.txt” ..... “cluster15.txt”.
4. Each file starts with three keywords (Note: ignore the character ‘u’ before each keyword)
5. The keywords look like this: [u‘abcd’, u‘pqrs’, u‘xyz’] and the text start from the next line

**Notes:**

1. **Caution:** Each cluster (file) contains a lot of data. This is important for my work.
2. Ignore “cluster11.txt” file. It contains some instructions for the students from the course material.
3. Also, each file may contain some instructions for students (like instructions for quiz, etc.). This may be ignored too.

**Evaluation:**

1. I would appreciate it if you could give us a brief approach as to how you evaluated the data.
2. How you read the text and how did you try to match the keywords with the text?
3. Do you think the keywords represent a good overview of the text? Do you think the keywords represent the skills associated with the text?