Exercise 5: Clustering and Class LMs

You can earn up to 10 points on this exercise.
You may work as a group of up to 3 people, but please submit your own version.
You may use any programming language you wish, but any submission that we cannot run on our computers without installing things must be presented to the class.

Please email your solution to `claytong@coli.uni-saarland.de` *or submit before the tutorial by* **10:14:59 AM GMT+1, December 2, 2015**.

Task 1

Consider the following five two-dimensional data points:

(−1,1)
(5,5)
(2,5)
(−5,5)
(2,1)

Illustrate hierarchical agglomerative clustering using standard Euclidean distance as your comparison metric. Explicitly state all assumptions that you make. (2 points)

Task 2

Suppose the following word classes:

preposition	*noun*	*verb*	*adjective*
of	they	are	the
with	men	read	some
near	telescopes	see	ugly

(a) Specify the parameters and values needed to fully define a class-based, bigram language model for these classes. For example "$p(X) = 0.2$", where $p(X)$ is a parameter that you define. Include start and end sentence symbols. You do not need to consult explicit data. Just use your own intuitions about these words or their analogues in your native language. (4 points)

(b) Qualitatively evaluate your language model. For example,

- What kinds of common and uncommon sentences can it predict well?
- What kinds of common and uncommon sentences cannot be predicted well?
- How good are the assumptions that the model makes?

(4 points)