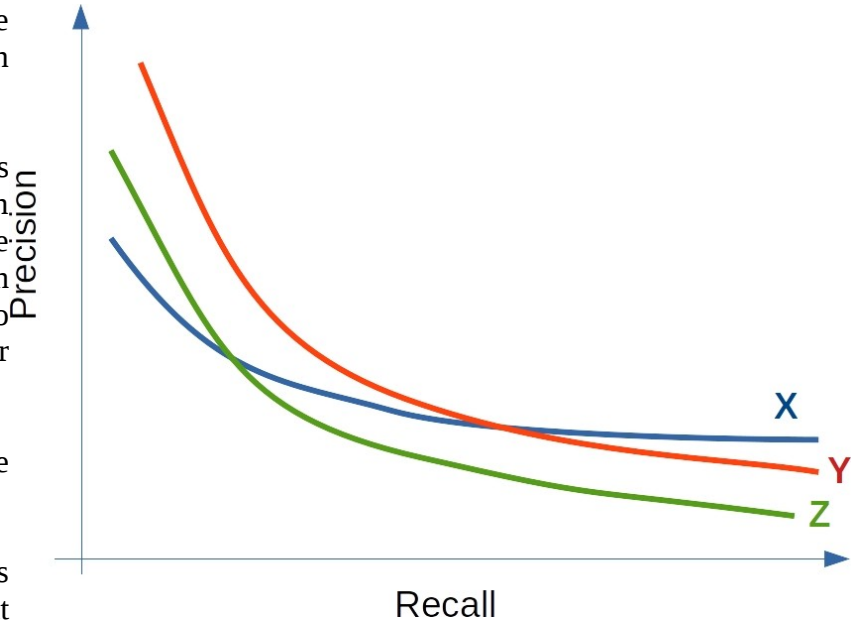
Web information retrieval – 2020/2021
Proff. Becchetti and Vitaletti

Exam – April 21th, 2021

Time: 60 minutes

Assignment 1:

1.1. Describe the procedure to obtain the picture on the right comparing search engines X(blue), Y(red) and Z(green).



1.2. The company LAW, providing services to lawyers, asks you to offer them a search engine where is crucial to provide all the documents related to a specific search query in order to allow their customers to consider all possibilities to help their assisted.

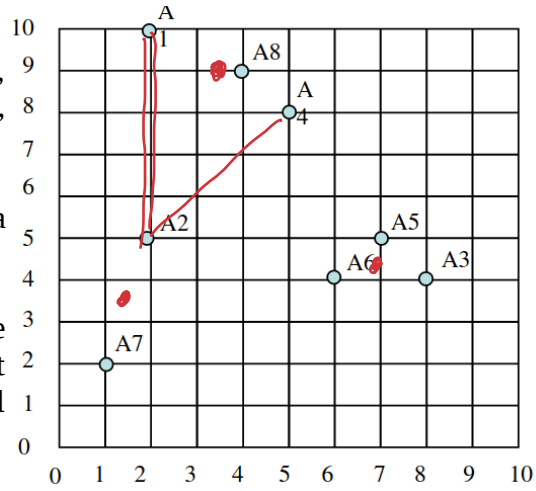
What is the system in the picture more suitable for LAW? Motivate the answer.

1.3. The company FOCUS instead, is interested in providing only documents that are actually relevant for a query.

What is the system in the picture more suitable for FOCUS? *You should motivate your answer*.

Assignment 2:

You have to cluster the points in the picture A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).



2.1. Your preliminary analysis shows that **3 clusters** are a good choice.

What kind of clustering algorithm would you use? Motivate the answer and show the resulting clusters and any relevant information on them. In particular discuss how some initial assumptions can assign a point to different clusters.

2.2. After a while you are informed that A1, A2 and A4 belong to the same class RED, and A5 and A6 to a different class BLUE. You don't have any other information that suggests you there are more than 2 classes. So you change your initial assumption and you assume now there are only 2 classes.

What kind of algorithm would you use in this case?
Motivate the answer and show the resulting clusters and any relevant information on them.

Assignment 3.
Consider the following set of documents:

D1. *He moved from London, Ontario, to London, England.*
D2. *He moved from London, England, to London, Ontario.*
D3. *He moved from England to London, Ontario.*

3.1. Which of the above documents have identical and different bag of words representations for the Bernoulli model?
3.2. Which of the above documents have identical and different bag of words representations for the multinomial model? If there are differences, describe them.
3.3. Supposing the above documents belong to the same class *c*, what is P(t|*c*) according to the Bernoulli model, if t = "London"?