Hybrid Email classifier: The Plan

1. Text Pre processing
	1. Tokenization using Stanford Parser.(Need to decide which parts of speech to be taken)
	2. Stop word removal. : to remove words such as “a”, “the”, “I”, “he”, “she”, “is”, “are”, etc
	3. Normalize words. : Stanford Lemma
	4. Then make instance based on top K term frequency.
	5. N grams
2. Training
	1. Design NBTree – Hybrid Decision Tree Naïve baiyes classifier.
	2. Train input on this, along with a Test Instance.
	3. Predict the category.

Preparing the data for Text categorization:

1. Find the training data set.
	1. For Spam : CSDMC2010 SPAM corpus
		* 4327 messages out of which there are 2949 non-spam messages (HAM) and 1378 spam messages (SPAM).
		* This Data set comes with a python script, to extract the contents. But I preferred writing my own Java code to read the contents.
	2. For Rest categories { Atheism,Medical,Autos,Sports }
		* 20 newsgroup data-set : <http://www.csmining.org/index.php/id-20-newsgroups.html>.
		* Atheism – 480 instances.
		* Medical – 594 instances.
		* Autos – 1192 instances.
		* Sports – 1180 instances.
2. Prepare the Training data:
	1. For Spam : Put all the 1378 SPAM classified instances of CSDMC2010 SPAM corpus into folder “data\train\spam”.
		* FIlterSpam.py, prepared a script that filters the SPAM training instances
	2. For Ham: Put all the 2949 SPAM classified instances of CSDMC2010 SPAM corpus into folder “data\train\ham”.
		* FIlterHam.py , prepared a script that filters the SPAM training instances
	3. For Atheism: Put 480 instances in “data\train\atheism”
	4. For Medical: Put 594 instances in “data\train\medical”
	5. For Autos: Put 1192 instances in “data\train\autos”
	6. For Sports: Put 1180 instances in “data\train\sports”