CS547: Information Retrieval **Spring 2016**

Homework #1

Due date & time: Should be submitted on blackboard by 7:00pmEST on Friday, January 29, 2016.

Late Policy: Late work will be penalized 10% per day (24 hour period). This penalty will apply except in case of documented emergency (e.g., medical emergency), or by prior arrangement if doing the work in advance is impossible due to fault of the instructor (e.g., you are going to a conference and ask to start the project early, but I don't have it ready yet.)

Additional Instructions: The submitted homework must be typed. Using Latex is highly recommended.

Problem 1 (10 pts) Describe an information retrieval problem that is NOT ad-hoc retrieval.

Problem 2 (10 pts) We often preprocess based on the structural properties of a document representation before indexing. Give an example of a query that would likely give very poor results if we used all words in an html document instead of preprocessing based on *structural* properties.

Problem 3 (24 pts) Text representation: Vocabulary

For each of the following, say if you would expect it to improve *precision, recall, both, or neither*. For each, give an example query and describe why it supports your answer.

1. Controlled Vocabulary
2. Stopwords
3. Stemming
4. Representing phrases as a single term.

Problem 4 (10 pts) Evaluation: F-measure

The *balanced F-measure* is defined as the mean of precision and recall. Explain why the balanced F-measure might not be a good "single number" to evaluate how useful an information retrieval system would be to an end user.

Problem 5 (20 pts) Vector Space Model

Given the following vector space model representation of a corpus:

	Doc1	Doc2	Doc3	Doc4
Purdue	4	0	2	1
Information	3	0	0	3
Retrieval	1	3	3	3

1. Give an inverted index representation of the corpus.
2. Compute the TF-IDF cosine similarity score between Doc3 and the query "Purdue Information". State any assumptions you make.
3. Which document is most likely about this course?