Blogosphere, an online community which encompasses all blogs, has grown rapidly in the last 6 years or so and various researches have been conducted to examine this special space. The WIDIT TREC team participated in the Blog track of 2006 Text Retrieval Conference (TREC) to investigate effective methods of retrieving blog posts that express an opinion about a given target.

1. Data
The task collection is a big sample of the blogosphere, including 3,215,171 permalink documents and spanning over a period of 11 weeks (December 05-February 06). There are 50 test topics which specified the retrieval targets. Since Blog track is a new track in 2006, there is no training data provided.

2. Research Questions
- Does noise reduction (e.g., exclusion of non-English blogs, navigational text, etc.) affect blog retrieval?
- What are the evidences of subjectiveness/opinion and how can they be leveraged to judge whether a permalink document is opinionated or not?

3. Methods
First, the collection is pre-processed and indexed; then, traditional IR approach is applied to retrieve on-topic documents; next, an opinion module is applied to the on-topic documents to compute subjectiveness scores; finally, on-topic results are re-ranked by the fusion score of the subjectiveness to boost the ranks of opinion documents to the top.

20 pseudo topics and about 700 training documents are built manually for tuning our opinion module and the entire system.

Our opinion module consists of four sub modules,

3.1 Subjective Terms Module
The basic idea is to identify a set of terms that only occur frequently in the subjective/opinion blogs and use these terms to score the subjectiveness of blogs. We first generate a candidate term list by ranking terms from the positive training data by information gain, selecting terms above a threshold, and removing terms that also appear in the negative training data. Next, we examine the candidate list to produce the final subjective term list.

3.2 Rare Terms Module
The hypothesis is that people become creative when they are expressing their opinion about something and will tend to use uncommon terms (e.g., “sooo good”). So we extract terms from positive training data, rank them by frequency and only select those terms with frequency smaller than 2. We then remove the dictionary terms from list and
examine the final rare term list to manually construct the regular expressions that will match rare term patterns used in opinion blogs.

**3.3 IU module**
IU (I and you) module comes from the observation that ‘I’ and ‘you’ appear in the opinionate blogs very frequently. The actual implementation is to find the collocation patterns of ‘I’, ‘you’, ‘we’, ‘your’, ‘me’, ‘our’, ‘my’ in the boundary of one sentence. Examples of this kind of co-occurrence are ‘I… your’, ‘I…my…you’, and ‘I…our’. Occurrences of such syntax structures are counted to score the IU opinion score.

**3.4 Adj/Verb module**
The hypothesis is that the more subjective verbs and adjectives are used in a document, the more likely is that the document is an opinionated one. We generate a subjective adjective-verb lexicon via a hybrid approach that involves different resources and technologies from computational linguistics, information retrieval and library science, and use the lexicon to score the subjectiveness of a document.

**4. Results and Evaluation**
We will get the relevance judgment at the beginning of October, by when we can conduct evaluation for out test run.