Here you will find deliverables of the projects done by members of WING, exclusive of publications. If you're looking for demonstrations of systems, they are listed with each project. WING members, if you are asked to add something to this list, please do, but please add at the top of each list and add your name as the person who compiled or coded the resource.

**Tools**

These are some of the in-house NLP and IR tools that we have built to facilitate our research at WING. We hope you'll find some of these tools helpful. A full list of all such tools that we have installed for research at NUS (including tons of ones from external sources can be found on our resource page). You may want to subscribe to our software-announce mailing list to be informed of new releases.

**A PDTB-Styled End-to-End Discourse Parser** - A discourse parser that outputs discourse relations in the Penn Discourse Treebank (PDTB) style -- deprecated: see below. The original source is still available [here](#). Annotates implicit/explicit relations, argument spans and attribution spans. Free for academic and research purposes only. Described in the technical report above. Coded by Ziheng Lin. Joint work with the NLP group led by A/P Ng Hwee Tou.

A newer version of the discourse parser was released in 2015 by our own Ilija Ilievski. He has recoded the discourse parser from the ground up and incorporated dependencies to make it much easier to run. It is completely compatible with the existing parser and is now the currently supported version. See it in Github: [https://github.com/WING-NUS/pdtb-parser](https://github.com/WING-NUS/pdtb-parser). Also joint work with the NLP group.

**CoNMF (Co-regularized Non-negative Matrix Factorization) dataset and code**. This extends NMF for multi-view clustering (MVC) by jointly factorizing the multiple matrices through co-regularization. Experimental results on Last.fm and Yelp comments datasets demonstrate the effectiveness of our solution. In Last.fm, CoNMF achieves accuracy 51.9%, better k-means with increase of 12%, while being comparable with the state-of-the-art MVC method Co-Spectral Clustering (CoSC). On a Yelp dataset, CoNMF achieves accuracy 67.6%, outperforms CoSC with performance gain of 7%. Coded by Xiangnan He.

**WING.NUS Keyphrase** - This is an open-source keyphrase generator, expressly created for
scholarly documents. Featured in SemEval2, Task 5 (Placed #2 out of 19 teams). Coded by Emma Nguyen, with input from Minh-Thang Luong.

**DiCE Tooltip Translator** - This is an open-source translation tooltip utility developed by the CSIDM undergraduates supervised in WING, for the Firefox browser. It translates from Chinese to English and vice versa and we are working utilities to enable word sense disambiguation and word segmentation.

**RAZ** - Robust Argumentative Zoning. A collaborative project between WING's Min and Simone Teufeul at the University of Cambridge.

**DefMiner** - This work extracts parts of definitions from scholarly articles using a conditional random field method. It is based on Yiping Jin's undergraduate research thesis and was published in EMNLP 2013.

**Kairos** - A scholarly paper crawling engine. Uses Lucene as a base and two CRF-based information extraction engines to run the process of gathering scholarly paper metadata. Coded by Markus Häense.

**Prastava** - An open source, ruby based recommendation system. Capable of using either collaborative filtering or context based filtering (standard IR methods), or a hybrid of both. Consists of a server and client pair for production recommendation systems. Can take input from flat files or from a connection to a database engine. Coded by Himanshu Gahlot and Tarun Kumar.

**ParsCit** - An open source CRF-based reference string parser and section labeler. Used by a number of digital library development groups worldwide. Works with references in English, Italian and French and over citations in the computer science, medical and humanities domains. Incorporates the Enlil author-affiliation matching algorithm library. Also available as a web service. Coded by Huy Do Nhat Hoang, Muthu Kumar C, Min-Yen Kan, Minh-Thang Luong and the CiteSeerX team at Penn State University.

**QANUS** - This is an open-source framework upon which new QA systems can be rapidly and easily developed. QANUS is also shipped with complete QA modules and can function as a simple baseline for other QA systems. Coded by Jun Ping Ng.
**Search engine wrapper** - This package provides a Java wrapper framework for unifying programatic access to search engines. It contains an API as well as a command-line application. Coded by Yee Fan Tan.

**SSID** - The Student Submission Integrity Diagnosis System, developed initially by Jon Yan Horn Poon. A source code plagiarism checker. Work published to ITiCSE 2012 (Haifa, Israel). Code also published to [github](https://github).

**SWING** - The Summarizer from the Web IR / NLP Group (WING), is a modular, state-of-the-art automatic extractive text summarization system. It produces informative summaries from multiple topic related documents using a supervised learning model. SWING is also the best performing summarizer at the international [TAC 2011](https://www.nist.gov/itl/iad/mig/tac) competition, getting high marks for the ROUGE evaluation measure. Coded by Jun-Ping Ng, Praveen Bysani and Ziheng Lin.

**Record Matching Package** - This record matching package is written as an extensible framework in Java, with the goal of making the writing of programs that perform record matching tasks easier. The focus here is on pairwise comparison of records, and this package includes building blocks for similarity or distance metrics, blocking algorithms, and clustering algorithms. A couple of metrics and algorithms are included in this package, and new metrics and algorithms can be implemented by subclassing suitable classes in this framework. Coded by Yee Fan Tan.

**JavaRAP** - A Java open-source reimplementation of the famed RAP (Resolution of Anaphora Program) by Boguraev and Kennedy. Note: this program is not considered competitive for anaphora resolution by today's standards but we have implemented it for benchmarking purposes. Feel free to download and use for non-commercial purposes. Coded by Long Qiu.

**Daemonized Collins parser** - Got more than a few sentences to parse? The Collins head driven parser is still considered one of the best open-source English language parsers. We've taken Michael's source code and wrapped it into a daemonized version that you can send sentences to through a socket service, avoiding the long initialization needed by the parser. Implemented by Min-Yen Kan.
Corpora and Datasets

These are text and image and other datasets used by experiments in our group. Most are freely available for research use (not commercial use in some cases).

**About.me dataset**. This dataset comprises of 15K users who have cross-linked accounts on four of six major public online social network accounts -- Flickr, Google+, Instagram, Tumblr, Twitter and Youtube. This dataset is the basis for our ASONAM 2015 paper "#mytweet via Instagram: Exploring User Behaviour across Multiple Social Networks". Collected by Dongyuan Lu and Bang Hui Lim; organization work into DB also assisted by Jerome Cheng.

**CoNMF (Co-regularized Non-negative Matrix Factorization) dataset and code**. This extends NMF for multi-view clustering (MVC) by jointly factorizing the multiple matrices through co-regularization. Experimental results on Last.fm and Yelp comments datasets demonstrate the effectiveness of our solution. In Last.fm, CoNMF achieves accuracy 51.9%, betters k-means with increase of 12%, while being comparable with the state-of-the-art MVC method Co-Spectral Clustering (CoSC). On a Yelp dataset, CoNMF achieves accuracy 67.6%, outperforms CoSC with performance gain of 7%. Coded by Xiangnan He.

**SciSumm Pilot Corpus**. This dataset consists of 10 ACL Anthology papers and the neighborhood of papers that cite them. The annotations indicate the provenance (source) of a citing sentence's claim in the original paper and what facet it belongs to (from a fixed set) -- following the annotation guidelines for summarization set out by the TAC BioMedSumm task. You can find the corpus and more details on GitHub: [https://github.com/WING-NUS/scisumm-corpus](https://github.com/WING-NUS/scisumm-corpus). A joint initiative by Muthu Kumar and research assistant alumnus Ankur Khanna in collaboration with Dr Kokil Jaidka, alumnus of Nanyang Technological University (NTU).

**Term & Definition Mining Corpus**: This corpus contains one manually annotated and one automatically extracted dataset. Both are derived from ACL Anthology Corpus and contain definitional sentences with definiendum (term) and definiens (definition) marked out. - Compiled by Yiping Jin.

**Weibo Chinese Word Segmentation and Informal Word Recognition Corpus** : This is a corpus
of 5,500 Weibo microblog posts sampled from PReV, with annotations done in Aobo’s ACL 2013 paper, that describes techniques to jointly recognize informal words and perform word segmentation for Chinese. Compiled by Aobo Wang.

**Enlil Document Collection**: Gold standard data for author-affiliation matching. This dataset is described in the Hoang et al. 2013 JCDL paper, entitled "Extracting and Matching Authors and Affiliations in Scholarly Documents". Compiled by Muthu Kumar and Huy Do.

**Chaptrs Personal Photograph Dataset** - This is the dataset described in Jesse's 2013 short JCDL paper entitled "Constructing an Anonymous Dataset From the Personal Digital Photo Libraries of Mac App Store Users", comprising of data of over 470K photos in 60K photosets. Compiled by Jesse Prabawa Gozali.

**Math Webpage Corpus with Readability Judgments** - This corpus contains 120 math webpages with readability annotation corresponding to 7 different education levels ranging from primary school to university. Compiled by Jin Zhao.

**Related Work Summarization Dataset** - This dataset comprises of 20 articles in the NLP / IR domain that have the texts of the papers as well as their referenced papers, and a manually-built hierarchical topic tree. The original related work section has also been set aside for evaluation purposes. Compiled by Cong Duy Vu Hoang.

**Scholarly Paper Recommendation Dataset** - This dataset contains gold-standard labels of 28 researchers in the NLP and IR areas, over NLP papers available in the ACL ARC dataset. Full citation and reference information about candidates papers as well as their computed feature vectors are included. Compiled by Kazunari Sugiyama.

**ACL Anthology Reference Corpus (ACL ARC)** - This dataset contains computational linguistics scholarly articles taken from a Feb 2007 snapshot of the ACL Anthology. The corpus contains PDFs, page images (in .png format), as well as the running text of papers, extracted by OCR and pdfbox. Compiled a consortium of university groups worldwide. Our group participated and
coordinated the development.

**Keyphrase Corpus** - The corpus consists of more than 200 scientific publications, each has 4 different formats: PDF, HTML, plain text, and XML. Compiled by Emma Thuy Dung Nguyen.

**NUS Scenario Corpus** - This corpus contains news articles for more than 15 scenarios. The goal was to collect 10 events, each represented by at least 5 article instances, for each scenario. The articles were taken from a controlled list of websites that 1) are true, online versions of articles provided by reputable news agencies, and 2) must provide free news article archives, extending back no less than 5 years. Compiled by Long Qiu.

**NUS SMS Corpus** - This is a corpus of about 80K Short Message Service messages from mobile phone users in Singapore and the world (collected by crowdsourcing). All in English or Mandarin Chinese. The contributors were mostly university students or crowdsourced workers who contributed messages for a small amount of renumeration. Compiled by Tao Chen and previously by Yijue How.

**Presentation to Document Alignment Corpus** - This is a manual alignment of 20 scholarly papers from the database community to their corresponding presentations. The alignments are from one slide to multiple paragraphs. Compiled by Eugene Ezekiel.

**Light Verb / Support Verb Annotations** - This is a corpus of light verb annotations (aka support verbs; e.g., "make a call") that were annotated to support a supervised learning algorithm to differentiate them from meaning bearing (heavy) verbs. Compiled by Yee Fan Tan.

**Javascript Functionality Annotations** [currently down] - Over 1.8K different JavaScript units have been extracted and annotated from the WT10G standard web corpus. These are all the unique JavaScript units that we were able to detect in the entire WT10G, although there were many duplicates in the original 20+K unit instances. Compiled by Wei Lu.

**NPIC Image Corpus** - This is a 4.7GB image collection, comprising of two different
collections: a spidered portion gathered from the Web and another portion taken from the freely-accessible Wikipedia Commons. Compiled by Fei Wang.

**Event-Timex Temporal Relationship Crowdsourced Corpus** - This corpus consists of 8,576 annotations of the temporal relationship between an event and time expression pair within a sentence. It is compiled via crowdsourcing, and has been demonstrated to be of comparable quality to a expert-curated corpus. Compiled by Jun-Ping Ng.