

# COMP90042 Web Search and Text Analysis

<b>Credit Points:</b>	12.5									
<b>Level:</b>	9 (Graduate/Postgraduate)									
<b>Dates &amp; Locations:</b>	2015, Parkville This subject commences in the following study period/s: Semester 1, Parkville - Taught on campus.									
<b>Time Commitment:</b>	Contact Hours: 36 hours, comprising of one 2-hour lecture and one 1-hour workshop per week Total Time Commitment: 200 hours									
<b>Prerequisites:</b>	<p><b>One of the following:</b></p> <table border="1"> <thead> <tr> <th>Subject</th> <th>Study Period Commencement:</th> <th>Credit Points:</th> </tr> </thead> <tbody> <tr> <td>COMP30018 Knowledge Technologies</td> <td>Semester 1, Semester 2</td> <td>12.50</td> </tr> <tr> <td>COMP90049 Knowledge Technologies</td> <td>Semester 1, Semester 2</td> <td>12.50</td> </tr> </tbody> </table>	Subject	Study Period Commencement:	Credit Points:	COMP30018 Knowledge Technologies	Semester 1, Semester 2	12.50	COMP90049 Knowledge Technologies	Semester 1, Semester 2	12.50
Subject	Study Period Commencement:	Credit Points:								
COMP30018 Knowledge Technologies	Semester 1, Semester 2	12.50								
COMP90049 Knowledge Technologies	Semester 1, Semester 2	12.50								
<b>Corequisites:</b>	None									
<b>Recommended Background Knowledge:</b>	None									
<b>Non Allowed Subjects:</b>	433-460 Human Language Technology 433-467 Text and Document Management 433-660 Human Language Technology 433-667 Text and Document Management 433-476 Text and Document Management									
<b>Core Participation Requirements:</b>	<p>&lt;p&gt;For the purposes of considering request for Reasonable Adjustments under the Disability Standards for Education (Cwth 2005), and Student Support and Engagement Policy, academic requirements for this subject are articulated in the Subject Overview, Learning Outcomes, Assessment and Generic Skills sections of this entry.&lt;/p&gt; &lt;p&gt;It is University policy to take all reasonable steps to minimise the impact of disability upon academic study, and reasonable adjustments will be made to enhance a student's participation in the University's programs. Students who feel their disability may impact on meeting the requirements of this subject are encouraged to discuss this matter with a Faculty Student Adviser and Student Equity and Disability Support: &lt;a href="http://services.unimelb.edu.au/disability"&gt;http://services.unimelb.edu.au/disability&lt;/a&gt;&lt;/p&gt;</p>									
<b>Coordinator:</b>	Assoc Prof Steven Bird									
<b>Contact:</b>	email: <a href="mailto:sbird@unimelb.edu.au">sbird@unimelb.edu.au</a> ( <a href="mailto:sbird@unimelb.edu.au">mailto:sbird@unimelb.edu.au</a> )									
<b>Subject Overview:</b>	<p><b>AIMS</b></p> <p>The aims for this subject is for students to develop an understanding of the main algorithms used in natural language processing and text retrieval, for use in a diverse range of applications including search engines, cross-language information retrieval, machine translation, text mining, question answering, summarisation, and grammar correction. Topics to be covered include text normalisation, sentence boundary detection, part-of-speech tagging, n-gram language modelling, and text classification. The programming language used is Python.</p> <p><b>INDICATIVE CONTENT</b></p> <p>Topics covered will include:</p> <ul style="list-style-type: none"> <li># Document classification, including gender detection, topic detection and language identification</li> </ul>									

	<ul style="list-style-type: none"> <li># Weighted finite state transducers, hidden Markov models</li> <li># N-gram language modelling, including statistical estimation</li> <li># Sentence segmentation and alignment, the IBM models, expectation maximisation</li> <li># Search algorithms including beam search, A* search</li> <li># Term indexing, vector space model, term weighting.</li> </ul>
<b>Learning Outcomes:</b>	<p><b>NTENDED LEARNING OUTCOMES (ILO)</b></p> <p>On completion of this subject the student is expected to:</p> <ol style="list-style-type: none"> <li>1 Articulate issues relevant to the efficient implementation of language processing systems and text retrieval systems</li> <li>2 Apply natural language processing and information retrieval methodologies to textual data</li> <li>3 Develop and evaluate computational models of language, based on results from the research literature</li> </ol>
<b>Assessment:</b>	<p>Project assignments will be done during the semester, requiring approximately 50 - 55 hours of work in total (40%). There are two projects, due around week 6 and week 12 A research-oriented workshop presentation, requiring approximately 13 - 15 hours of work (10%) One 2-hour end-of-semester examination (50%). Hurdle requirement: To pass the subject, students must obtain at least: 50% overall 25/50 in the continuous assessment 25/50 in the end-of-semester written examination. Intended Learning Outcomes (ILOs) 1 and 2 are addressed in the lectures, workshops, and exam; ILOs 3 and 4 are addressed in the project work and oral presentation.</p>
<b>Prescribed Texts:</b>	None
<b>Breadth Options:</b>	This subject is not available as a breadth subject.
<b>Fees Information:</b>	Subject EFTSL, Level, Discipline & Census Date, <a href="http://enrolment.unimelb.edu.au/fees">http://enrolment.unimelb.edu.au/fees</a>
<b>Generic Skills:</b>	<p>On completing this subject, students should have the following skills:</p> <ul style="list-style-type: none"> <li># Formulate and implement algorithmic solutions to computational problems, with reference to the research literature</li> <li># Apply a systems approach to complex problems, and design for operational efficiency</li> <li># Design, implement and test programs for small and medium size problems in the Python programming language.</li> </ul>
<b>Notes:</b>	<p><b>LEARNING AND TEACHING METHODS</b></p> <p>The subject comprises a weekly 2 hour lecture followed by a 1 hour laboratory exercise. Weekly readings are assigned from the research literature, and weekly laboratory exercises are assigned. Additionally, a significant amount of project work is assigned.</p> <p><b>INDICATIVE KEY LEARNING RESOURCES</b></p> <p>At the beginning of the semester, the coordinator will post a list of readings from the research literature and research monographs which will form the basis of the intellectual content of the subject. An indicative monograph is <i>Statistical Machine Translation</i>, by Philipp Koehn (2010).</p> <p><b>CAREERS / INDUSTRY LINKS</b></p> <p>A growing sector of the IT industry is concerned with leveraging the information that is locked up in semi-structured text data on the web. Large scale analysis and exploitation of this information depends on graduates with a solid grounding in natural language processing and text retrieval algorithms, and experience with implementing systems that are informed by the research literature.</p>
<b>Related Course(s):</b>	<p>Master of Information Technology  Master of Philosophy - Engineering  Master of Science (Computer Science)  Master of Software Systems Engineering</p>

	Ph.D.- Engineering
<b>Related Majors/Minors/ Specialisations:</b>	B-ENG Software Engineering stream Computer Science Computer Science MIT Computing Specialisation MIT Distributed Computing Specialisation Master of Engineering (Software)