

Sanaindeksi

Käyttöohje

Patrik Ahvenainen*
Helsingin yliopisto
Fysiikan laitos

1. syyskuuta 2012

Tämä käyttöohje koskee vain **pysanaindeksi**-pääohjelmaa. Ohjelma käännetään ja ajetaan python-tulkilla komentoriviltä komennolla `python pysanaindeksi.py`. Ohjelma on suunniteltu kieliversiolle 2.6 eikä se toimi ainakaan python 3.x kieliversioilla. Pääohjelma löytyy `src/WordIndex`-alihakemistosta.

Indeksoitavat tiedostot täytyy siirtää Materials-alihakemistoon. Tiedostojen fonttikoodauksen tulisi olla UTF-8. Tiedostossa saa olla mitä vain merkkejä, mutta vain merkit "A-Z", "0-9", "'" ja "-" hyväksytään. Kirjaimet hyväksytään myös pieninä, jolloin ne muunnetaan isoiksi ennen hakuja. Vain hyväksytyillä merkeillä alkavat sanat indeksoidaan ja sana katkaistaan ensimmäiseen ei-hyväksytyyn merkkiin.

Pääohjelmassa navigoidaan pääasiassa valitsemalla numerovalinnalla haluttu toimenpide tai vaihtoehto. Muita kuin kokonaislukuja saa syöttää vain haun yhteydessä.

Haussa saa käyttää vain hyväksytyjä merkkejä (muita ei oteta huomioon, samojen sääntöjen mukaan kuin indeksoinnissakin). Ainut poikkeus tästä on sulut "(" ja ")". Näitä voidaan käyttää ryhmittelemään hakusanoja ja niiden täytyy esiintyä omana sananaan (välilyönnein eroteltuna muusta hakukomennosta). Erityissanoja ovat AND, OR, NOT ja XOR. Näistä AND etsii sekä vasemmalla, että oikealle esiintyviä termejä, OR vähintään yhtä termeistä, XOR vain toista termeistä ja NOT rivejä, joissa esiintyy vasemman puolen termi, muttei oikean puolen termi. AND-erityissanon käyttö ei ole välttämätöntä, sillä termien, joiden välissä ei ole mitään erityissanaa, väliin oletetaan AND. Myös erikoismerkkiä * voidaan käyttää sanan katkaisuun; jos sitä etsitään löydetään sanat, jotka alkavat sitä välittömästi edeltävillä merkeillä. Hyviä ja huonoja hakutermejä on esitetty taulukoissa 1 ja 2.

*patrik.ahvenainen@alumni.helsinki.fi / 013326292

Taulukko 1: Hyväksyttäviä hakutermejä. Samalla rivillä esiintyvät haut ovat indenttisiä, eli tuottavat saman hakutuloksen.

cat	CAT	cAt
ca*	ca* OR cat	
ca* OR mous *		
cat AND mouse	cat mouse	(cat AND mouse)
horse OR car	HORSE OR CAR	(horse OR car)
(horse OR car) NOT apple		
cat OR mouse NOT (cat mouse)	(cat mouse) XOR (cat OR mouse)	cat XOR mouse

Taulukko 2: Huonoja hakutermejä. Joko täysin väärin tai harhaanjohtavia.

a and b	a or b	a Not b
□	äiti	°
(sana)	(sana OR lause)	(sana AND toinen) NOT kolmas
sana AND NOT toinen	XOR sana	AND