

# Aranea Web-Crawled Corpora: A Source of Diverse and Unified I language Data for NLP

Wladimir Benko [wladimir.benko@juls.savba.sk](mailto:wladimir.benko@juls.savba.sk)

Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia

Comenius University Science Park, UNESCO Chair in Plurlingual and Multicultural Communication, Bratislava, Slovakia

<http://aranea.juls.savba.sk/guest> | <http://unesco.uniba.sk/guest>

Search engine interface for København. Query: København, 30,211 > Random sample 100 (0.80 per million). Results include links to various news and information sites like highways-usa.com, cs.wikipedia.org, and others.

Search engine interface for Tbilisi. Query: Tbilisi, 97,301 > Random sample 100 (0.80 per million). Results include links to news and information sites like argnet.net, tbi.ge, and others.

Search engine interface for Tashkent. Query: Tashkent, 111,243 > Random sample 100 (0.80 per million). Results include links to news and information sites like ahol.tashkent.uz, uz.wikipedia.org, and others.

Search engine interface for København. Query: København, 30,211 > Random sample 100 (0.80 per million). Results include links to various news and information sites like highways-usa.com, cs.wikipedia.org, and others.

Search engine interface for Warszawa. Query: Warszawa, 45,723 > Random sample 100 (0.80 per million). Results include links to news and information sites like onet.pl, tvp.pl, and others.

Search engine interface for Moskva. Query: Moskva, 48,584 > Random sample 100 (0.80 per million). Results include links to news and information sites like rusnews.ru, newsru.com, and others.

Search engine interface for København. Query: København, 30,211 > Random sample 100 (0.80 per million). Results include links to various news and information sites like highways-usa.com, cs.wikipedia.org, and others.

Search engine interface for Warszawa. Query: Warszawa, 45,723 > Random sample 100 (0.80 per million). Results include links to news and information sites like onet.pl, tvp.pl, and others.

Search engine interface for Moskva. Query: Moskva, 48,584 > Random sample 100 (0.80 per million). Results include links to news and information sites like rusnews.ru, newsru.com, and others.

Search engine interface for Kazan. Query: Kazan, 215,280 > Random sample 100 (0.80 per million). Results include links to news and information sites like pravda.tatarstan.ru, kaza-taract.narod.ru, and others.

Search engine interface for Tehran. Query: Tehran, 66,190 > Random sample 100 (0.80 per million). Results include links to news and information sites like w.farhif.ir, arivash.ir, and others.

Search engine interface for Minsk. Query: Minsk, 65,317 > Random sample 100 (0.86 per million). Results include links to news and information sites like ecologies.by, minsk.by, and others.

Search engine interface for Kazan. Query: Kazan, 215,280 > Random sample 100 (0.80 per million). Results include links to news and information sites like pravda.tatarstan.ru, kaza-taract.narod.ru, and others.

Search engine interface for Tehran. Query: Tehran, 66,190 > Random sample 100 (0.80 per million). Results include links to news and information sites like w.farhif.ir, arivash.ir, and others.

Search engine interface for Minsk. Query: Minsk, 65,317 > Random sample 100 (0.86 per million). Results include links to news and information sites like ecologies.by, minsk.by, and others.