Topics and collocates

Other corpus websites such as Sketch Engine and BNCWeb (and English-Corpora.org as well) allow you to find the collocates for a given word, where the collocates are words that co-occur in a "span" from 3 or 4 words to the left and 3-4 words to the right of a "node word".

But in addition to showing collocates, English-Corpora.org has the only corpora (that we're aware of) that allow you to find words (which we will call "topics") that co-occur *anywhere in the text*. For example, a text with the word *seafood* might also have the words *food, dining, dinner, chef, meal, beach, wine, shrimp* in the same text – but not necessarily right near the word *seafood*.

The following are 30 words from iWeb, with their collocates (within 4 words left/right) and topics (words that co-occur anywhere in the text/webpage). The words are color coded for noun, verb, adjective, and adverb. The number following the node word (e.g. .45 for *arthritis*) shows what percent of the topics are different than the collocates, and the words that are bolded are the ones that are different in collocates and topics.

The bottom line is that the topics (words that co-occur *anywhere* in the text) definitely provide great insight into the meaning of the node word. And topics are only available from English-Corpora.org.

Node word	**Collocates**	**Topics**
% not same	within 4 word left/right	anywhere in the text
arthritis .45	rheumatoid pain disease **psoriatic** **condition** patient joint symptom treatment **cause** **treat** **suffer** knee **inflammatory** **diabetes** **osteoarthritis**	pain **joint** disease joint **inflammation** symptom treatment rheumatoid patient **bone** **chronic** **medication** knee **tissue** **muscle** **doctor**
asthma .45	allergy **attack** symptom **asthma** disease severe patient **condition** chronic **suffer** treatment **diabetes** respiratory allergic medication **treat**	symptom allergy **lung** disease medication respiratory allergic patient chronic treatment **medicine** **doctor** severe **breathe** **infection** **immune**
bra .45	wear sport size strap fit cup panties **nursing** comfortable breast **strapless** **lace** **padded** **push-up** **underwear** **underwire**	wear breast size cup strap comfortable **dress** **fabric** **clothes** fit sport **woman** **fit** panties **boob** **pants**
Catholic .45	Roman church protestant **orthodox** **Irish** protestant **devout** holy **jew** Christian faith Christian **faithful** religious **eastern** **Jewish**	church **pope** faith **priest** religious Roman holy **bishop** Christian **religion** Christian protestant protestant **doctrine** **prayer** **bible**
flooring .45	wood hardwood **laminate** tile vinyl install carpet kitchen wall **oak** floor **engineered** **solid** **rubber** **plank** **wooden**	floor wood tile kitchen **room** **bedroom** hardwood carpet **bathroom** install wall vinyl **installation** **ceiling** **home** **dining**
volcano .45	**active** erupt eruption park earthquake island mountain **extinct** lava **dormant** **mount** **bay** crater lake **arenal** **ancient**	island **volcanic** eruption lava mountain **earth** **beach** lake **rock** **tour** **ocean** park crater **sea** erupt earthquake

Node word % not same	**Collocates** within 4 word left/right	**Topics** anywhere in the text
yacht .45	club charter **luxury** **royal** **motor** boat sail race **private** sailing race **super** **cruise** crew **broker** **ship**	boat sail sailing crew **island** club race **sea** charter **deck** **vessel** **beach** **guest** race **bay** **sailor**
condo .50	partment building unit bedroom rental beach **luxury** **sale** locate **owner** rent **association** **hotel** **tower** vacation **complex**	bedroom beach unit rental building **estate** apartment **kitchen** **pool** vacation **floor** rent **amenities** **property** **buyer** locate
magnesium .50	calcium potassium iron vitamin deficiency **zinc** mineral supplement **sulfate** **contain** **chloride** **alloy** **phosphorus** sodium **copper** **manganese**	calcium vitamin mineral **acid** supplement **nutrient** **diet** potassium **blood** deficiency **muscle** **salt** iron **protein** **healthy** sodium
melody .50	song chord lyrics rhythm **harmony** sing note guitar **vocal** **catchy** **maker** piano **beat** **sweet** instrument **arrangement**	song **music** **album** **musical** chord guitar sing piano **band** lyrics **tune** rhythm **sound** instrument **musician** note
seafood .50	fresh restaurant meat fish dish **steak** **eat** **sustainable** **vegetable** delicious **poultry** **chicken** **pasta** menu **salad** **catch**	restaurant fish dish fresh meat menu **beach** **food** **dining** **dinner** **chef** **meal** delicious **cook** **wine** **shrimp**
ballet .55	dancer theater royal opera dance **American** **national** class **flat** classical **jazz** perform dance **modern** **shoe** **contemporary**	dance dancer dance theater opera **music** **art** **performance** perform classical royal class **choreographer** **artistic** **production** **musical**
caffeine .55	**contain** alcohol coffee tea **amount** **effect** **intake** sugar **avoid** drink consume **stimulant** drink **mg** **consumption** **dose**	coffee drink tea drink **diet** **supplement** **cup** sugar **sleep** consume **energy** **ingredient** **acid** **fat** **vitamin** alcohol
chapel .55	funeral **Sistine** church **home** **wedding** memorial **royal** **lady** **pm** **ceiling** **hill** hall cemetery **choir** holy **dedicate**	church **century** funeral **cathedral** memorial holy **Catholic** cemetery hall **museum** **tower** **prayer** **building** **sister** **altar** **castle**
dinosaur .55	fossil bone earth bird **egg** museum species extinction **extinct** **roam** **skeleton** **giant** **prehistoric** **toy** **evolve** **footprint**	fossil museum **animal** species **creature** **scientist** earth bone **park** bird **kid** **evolution** **rock** **discovery** extinction **discover**
mortality .55	rate **high** **morbidity** infant **reduce** risk cancer **maternal** **cause** disease increased associate **all-cause** patient **reduction** **cardiovascular**	disease patient **death** **study** **population** risk **clinical** cancer **incidence** **health** rate **outcome** **infection** associate increased infant
recycling .55	**program** waste **center** facility plastic **bin** collection **centre** **electronics** recycle **waste** **household** container **reuse** **battery** **metal**	recycle waste plastic **landfill** **material** **environmental** collection container **green** **bag** facility **collect** **disposal** **trash** **recycled** **resident**

Node word % not same	**Collocates** within 4 word left/right	**Topics** anywhere in the text
resin .55	epoxy plastic **polyester** **cast** kit **mix** **composite** metal wood **acrylic** cure **fiberglass** layer coating **synthetic** **casting**	plastic **surface** **material** epoxy wood kit **paint** cure metal layer **paint** **coat** **dry** **glass** **mold** coating
honesty .60	integrity respect **appreciate** academic **transparency** **openness** truth trust **fairness** **brutal** **intellectual** **responsibility** **loyalty** **courage** trust **sincerity**	**honest** integrity truth respect **relationship** **lie** **moral** **conduct** trust **ethical** **lie** trust academic **cheat** **feeling** **dishonesty**
touchdown .60	yard score pass interception **throw** season **catch** **pass** **rushing** rush **career** **reception** **passing** **return** **carry** **run**	yard pass **quarterback** **football** season **receiver** **offense** **defense** **play** **quarter** interception **defensive** **bowl** **offensive** rush score
vampire .60	**diary** **slayer** werewolf blood **bat** **hunter** **weekend** **buffy** **lord** zombie human witch **killer** **ghost** novel **twilight**	**horror** **film** blood **monster** **movie** **episode** **character** **creature** **demon** **series** novel witch human **spell** werewolf zombie
kindness .65	act random love compassion **generosity** **respect** loving **treat** **stranger** **patience** **appreciate** mercy **empathy** **goodness** **spread** **gentleness**	act compassion love random **kind** **lord** **happiness** mercy **sin** **grace** **thank** **spiritual** loving **joy** **prayer** **blessing**
porch .65	front **back** **sit** **covered** deck **patio** **screened** **light** **screen** door **swing** **roof** **front** **entrance** floor **rear**	**bedroom** **kitchen** front **room** floor **house** **dining** **home** door **bath** **bathroom** **living** **outdoor** **fireplace** deck **garage**
scar .65	tissue **leave** acne skin heal **form** **fade** **mark** face **appearance** surgery **bear** **formation** **emotional** wound **visible**	skin tissue surgery heal acne **treatment** **pain** wound **infection** **surgeon** **healing** **procedure** **doctor** face **surgical** **painful**
telescope .65	**space** **radio** sky **optical** **binoculars** observatory observe **powerful** **mirror** astronomer **lens** **mount** solar **infrared** **eyepiece** **ground-based**	**star** astronomer **planet** **earth** sky **galaxy** **universe** observatory solar **sun** **astronomy** **observation** observe **object** **moon** **scientist**
throttle .65	**body** **full** **response** **position** **control** **cable** **open** engine **open** sensor **plate** valve **wide** speed **close** **pedal**	engine **fuel** **motor** valve **intake** **exhaust** **torque** **idle** sensor **gear** speed **brake** **bike** **car** **rear** **cylinder**
filmmaker .70	documentary film **independent** **artist** **photographer** **writer** actor **award-winning** direct producer **indie** **musician** festival **aspiring** **journalist** **acclaimed**	film **movie** documentary festival **cinema** actor **director** **audience** producer **scene** **shoot** **production** direct **screening** **camera** **footage**

Node word	Collocates	Topics
% not same	within 4 word left/right	anywhere in the text
obesity .70	diabetes **childhood** disease **rate** risk **epidemic** overweight **heart** weight **associate** **reduce** **cancer** **factor** **prevent** **cause** **link**	weight diabetes **diet** disease **obese** **healthy** **fat** overweight **health** risk **study** **fat** **food** **sugar** **eat** **researcher**
quilt .70	pattern block **top** fabric **baby** sew **patchwork** **finish** **shop** **bed** **cover** piece **pillow** square **blanket** **crazy**	fabric sew pattern block **sewing** **quilter** square **cut** **machine** **stitch** **seam** **stitch** piece **strip** **thread** **piece**
recession .75	**great** economic economy **global** **hit** **recent** **deep** **recovery** crisis unemployment **depression** **recover** **severe** **gum** **depth** inflation	economy economic **growth** unemployment **debt** inflation crisis **rate** **spending** **economist** **rise** **sector** **investor** **bank** **market** **labor**

(For those who are interested in how we find the topics / co-occurring words in the text/webpage).

1. For each of the texts in the corpus (22+ million web pages for iWeb, and ~500,000 texts for COCA), we found the words that occurred at least two times in the text/webpage, and which had a "normalized" frequency at least 20 times that of the entire corpus. This would typically eliminate most high frequency words or function words.

To give an example from iWeb (and things would be similar for COCA), the word *with* has a frequency of about 115,000,000 in the 14 billion words, or about 8,200 tokens per million words. Suppose that *with* occurs 9 times in a text with 1,000 words, giving a normalized frequency of 9,000 per million words. That is only about 1.1x the expected frequency -- much less that the 20x needed for the list of "keywords" from that text. On the other hand, take the word *telescope*. It has a frequency of about 120,000 tokens in the 14 billion words, or about 8.6 tokens per million words. If *telescope* occurs 3 times in a text with 1,000 words, that would be a normalized frequency of 3,000 tokens per million words. This is more than 200 times the expected frequency of 8.6, and so *telescope* is included as a "keyword" for that text.

2. For each of the "content" words (nouns, verbs, adjectives, and adverbs) in the 60,000 word list, we found which "keywords" (from #1) co-occur the most in the texts in the corpus (again, 22+ million web pages for iWeb and ~500,000 texts for COCA). In iWeb, for example, *star* is found as a keyword in 5,483 texts where *telescope* occurs, *galaxy* co-occurs 3,744 times, *universe* 3,036 times, and so on.

Because we use relational databases for the underlying architecture of all of the corpora from English-Corpora.org, we can do steps #1-2 very quickly – just about 20-30 minutes for even a corpus like iWeb – with 22+ million texts and 14 billion words total. It's hard to imagine how this could be done using any other architecture.

Finally, some people might say, "well, the reason that you find so many additional good words as "topics" is because your collocates are not that good; they are not done correctly. We have another page that discusses in detail how we do collocates, and compares our approach to other approaches, where there are multiple "fancy" statistic measures for collocates. Bottom line, we believe that our collocates are as good (probably much better) than those with other approaches.