Project 1: Words and Phrases

Md Mohaiminul Islam
Iowa State University
mmislam@iastate.edu

1 Methodology

We used three different experimental approaches to compare the performance of phrase extraction. Firstly, we used the naive greedy approach demonstrated in the tutorial in (Ganesan, 2019). This approach first splits text into coarse segments using punctuation, and then refines boundaries with stop words. Candidate phrases are generated across the corpus and ranked by frequency to assess quality. Minimum number of appearances to consider a phrase for tagging was set to $m = 10$ for our setup. We also maintained the same value of m for the other experiments as well to ensure benchmarking consistency.

Secondly, we conducted a statistical mixed method approach by started by taking all *N-grams* of length l, where $2 \leq l \leq 4$ and min frequency $m = 10$. Then we filter collocations based on *PMI*(Pointwise Mutual Information) to extract more quality phrases. A threshold for minimum *PMI* was set to 4. In parallel, a list of noun phrases were collected by *spaCy*'s linguistic analyzer and a separate set of top phrases were collected using TF-IDF scoring metric. Then all the sets were merged to create the combined list of phrases.

The third setup employs *BioBERT*'s (Lee et al., 2020; base cased v1.1, 2021) contextual embeddings to perform semantic similarity clustering of candidate phrases, identifying conceptually related multiword expressions through cosine similarity of their representation vectors. We also used *SciSpaCy* model, trained in biomedical corpora, for quality scoring of dense semantic phrases.

2 Main Results

2.1 Word cloud Visualizations

In Figure 1 Word cloud visualizations created from the cleaned and preprocessed corpus show certain key insights:

Observation: The raw frequency-based approach mainly highlights the most common terms such as '*trait*', '*QTL*', '*gene*', '*study*', '*SNP*' etc. and the **TF-IDF** based visualization highlights words like '*SNP*', '*gene*', '*QTL*', '*region*', '*association*' etc.

Insight: This shows TF-IDF down-weights common words such as '*trait*' and '*study*' most likely due to them being present in almost every document. Hence, their Inverse Document frequency was too low even if the raw frequency was high. If we look past the most dominating terms and observe the moderately dominating words, we can also see TF-IDF based image highlights words such as '*chicken*', '*milk*', '*sheep*', '*meat*', '*pig*', '*breed*' which the prior image does not. This reflects the words for identifying broad research topics within the corpus. This words are likely the most informative keywords to search with within the corpus.

2.2 Word2vec results

Some example outputs from Word2vec(Mikolov et al., 2013) model trained with the given parameters and a skip-gram algorithm :

Correct Examples : Some correct contextual results include: **QTL**→{**gga1**, **gga3**, **ssc1**$\cdots$}. These might be part of genomic sequences of different animals that carry qualitative traits. Also, for **Region** we have {**segment**, **vicinity**, **middle**$\cdots$} etc. words reflecting different genomic positions. This also makes sense in grammatical terms. Some other examples are **Chromosome**→{**centromeric**, **distal**, **autosomal**$\cdots$} in the context of similar biological concepts and **Analysis**→{**software**, **multivariate**, **regression**$\cdots$} etc. in the context of analytic techniques.

Negative/Questionable Examples: Most of the negative results are attributed to a bad job cleaning on the author's part and as a result some noise (numeric measurements, proper nouns and generic terms) was leftover in the model's vocabulary. For example, **SNP**→{**thirty**, **0.5**, **diego**· · ·} and **Study**→{**race**, **poorly**· · ·} etc.

Key Observations: Overall, the model seems to capture the context correctly in area-specific knowledge. It decently infers which words should be around a given word in different genomic, biomedical, or analytic contexts. Adding a custom stop word list to root out semantically less meaningful could improve the performance. The full list of words is attached in Appendix A

2.3 Phrase Mining

The findings of our phrase mining experiments are summarized in Table 1

Method	Phrases	Exact Matches
Naive	1762	80 (4.5%)
Multi method	1074	51 (4.7%)
Transformer based	800	6 (0.75%)

Table 1: Phrase mining methods validated against

Though the naive approach mined more phrases it was due to limits set on the second approach. AS the second method was a mixed method approach involving multiple extraction and refinement pipelines it was much slower. Therefore, number of phrases generated per method was capped at 500 per method and only the highest scoring phrases were selected per method. It shows a slight higher rate in terms of the phrases showing up in the '*Trait_dictionary.txt*' file. Although we can generalize that both methods were almost equal in that manner. But the third method could not generate many quality phrases. It could be an implementation error or a data cleaning mishap.

Retraining Word2vec: Retraining word2vec with phrase-tagged corpora yields clear differences. In this experiment, the second approach empirically produces better results. So, we discuss that in detail. We can observe the inclusion of many phrases in the similar word list for the top 10 TF-IDF words. The complete list is provided in Appendix A

Positive and Negative examples: We now have top words like **QTL** clustering with phrases like {'**QTL_trait**', '**QTL_chromosome**', '**fat_trait**'} etc. which are relevant phrases when it comes to identifying specific traits. We also see, **Gene**→{'**gene_snp**','**transcription_factor**', '**positional_functional_candidate_gene**'· · ·} and **Trait**→{'**affect_trait**', '**production_trait**', '**genetic_correlation**'· · ·} which are semantically much more meaningful with the given words rather than its constituencies. Negative examples are relatively few and far between compared to earlier. Some of them still being the noise of numerical values such as **Genotype**→{'**p<0.05**'}, while other examples have overmerged generic phrases such as **Region**→{'**region_identify**'}.

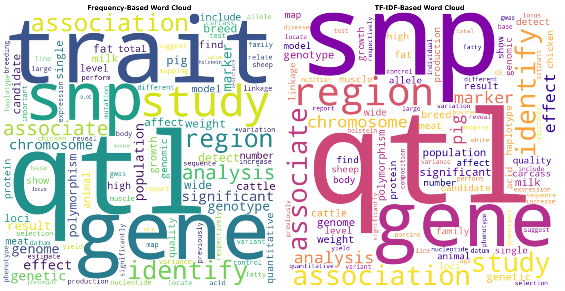

Figure 1: Comparison between word cloud visualizations based on raw frequency vs. using TF-IDF.

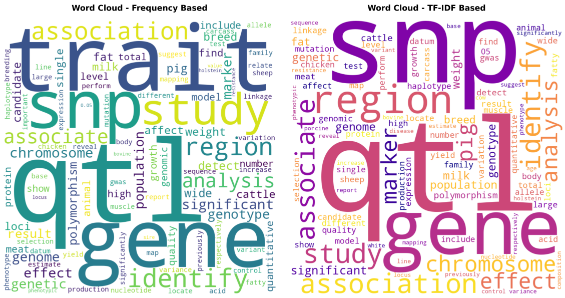

Figure 2: Word cloud visualizations after statistical multi method phrase tagging (method 2).

3 Discussion

Overall, for most key terms (QTL, gene, SNP, region), their neighbours look biologically more correct and useful after phrase mining. Cosine similarity scores have also increased across the board, which reflects that words are more tightly clustered in the vector space. This can be interpreted as the confidence of the model in its predictions. Specifically similar words that included noise and meaningless junk have reduced significantly. The first method also produces close results, but overall the second method is much more correct and meaningful in generating coherent predictions than the no-phrase baseline. We include the word cloud visualizations for method 1 and 3 in Appendix A.

References

BioBERT base cased v1.1. 2021. Huggingface repository.

Kavita Ganesan. 2019. How to incorporate phrases into word2vec – a text mining approach.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

A Appendix A

Figures 3, 4 are the word clouds generated after phrase tagging for methods 1 and 3.

Tables 2, 3, 4 Complete list of word2vec similar words with similarity scores for the no-phrase baseline, phrase tagging methods 1 and 2 respectively :

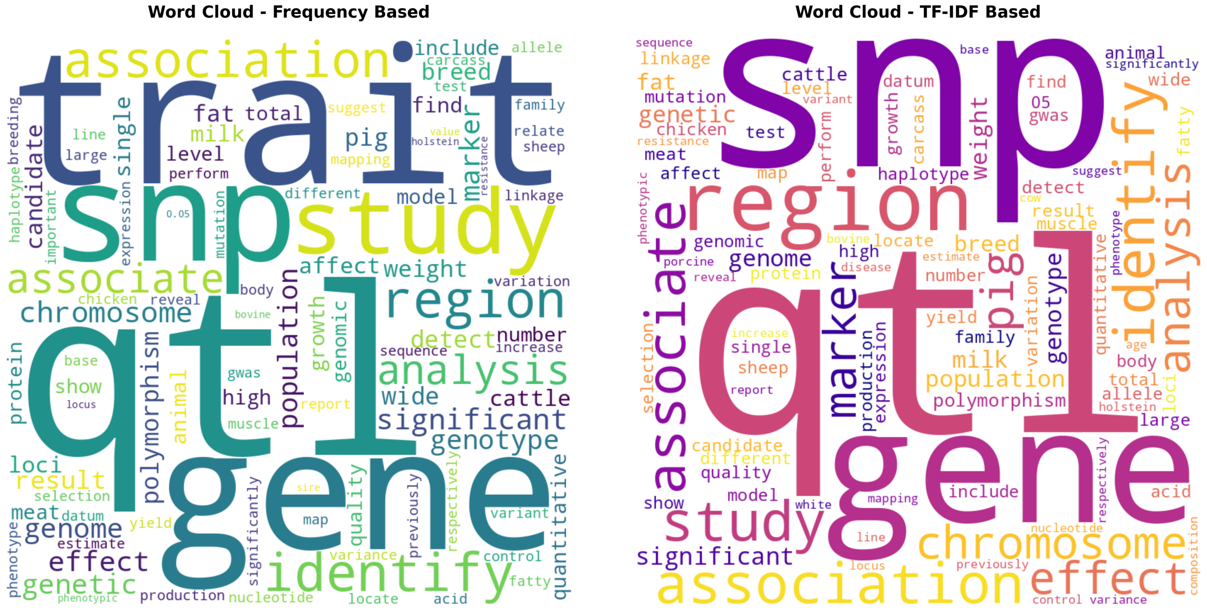

Figure 3: Word cloud visualizations after naive greedy phrase tagging.

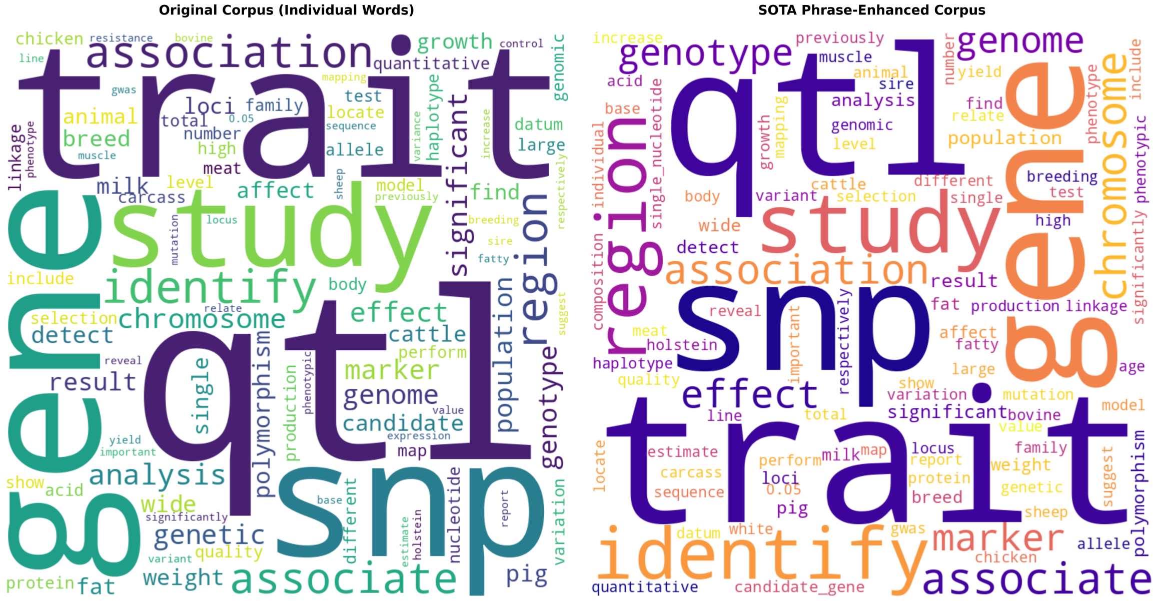

Figure 4: Word cloud visualizations after pretrained transformer based phrase tagging.

Query Word	**Top 20 Most Similar Words (similarity score)**
QTL	arm (0.672), qtls (0.662), localize (0.643), coincide (0.631), gga3 (0.617), harbour (0.614), chromosome (0.614), gga5 (0.614), endocrine (0.613), fat1 (0.600), imprinting (0.596), bta7 (0.594), presence (0.590), rhm (0.588), localization (0.586), distal (0.586), bta11 (0.586), sus (0.584), ssc9 (0.584), vicinity (0.583)
SNP	thirty (0.640), 129 (0.639), adjacent (0.638), bonferroni (0.610), intronic (0.605), infer (0.604), genbank (0.601), build (0.597), conditional (0.592), 0.5 (0.591), nucleotide (0.583), bovinesnp50 (0.582), glm (0.581), uncover (0.577), nearby (0.575), 119 (0.570), flank (0.569), bta5 (0.564), vicinity (0.564), bayes (0.564)
GENE	plausible (0.731), gnas (0.684), functional (0.683), positional (0.679), proximity (0.666), thyroid (0.662), igf1 (0.658), highlight (0.655), promote (0.646), spp1 (0.644), annotation (0.643), bioinformatics (0.643), literature (0.641), upstream (0.640), ontology (0.637), list (0.636), mirna (0.632), elovl6 (0.632), enrich (0.630), mitf (0.629)
REGION	segment (0.669), vicinity (0.660), mbp (0.654), proximal (0.644), harbour (0.639), middle (0.635), encompass (0.628), centromeric (0.627), harbor (0.618), sixteen (0.607), bta20 (0.606), overlap (0.604), nearby (0.600), locate (0.591), span (0.590), bta18 (0.585), bta11 (0.581), interestingly (0.576), oar3 (0.575), flanking (0.569)
ASSOCIATION	conditional (0.649), gwa (0.592), gemma (0.590), chi (0.585), weighted (0.579), nearby (0.573), bonferroni (0.568), genomewide (0.566), fst (0.564), package (0.563), bioinformatics (0.543), combination (0.539), methodology (0.535), retain (0.535), series (0.533), genbank (0.532), univariate (0.532), present (0.530), statistical (0.528), 129 (0.527)
IDENTIFY	detect (0.677), consistently (0.645), nearby (0.626), vicinity (0.589), locate (0.589), find (0.589), discover (0.584), kit (0.583), sixteen (0.578), harbour (0.577), mbp (0.576), identification (0.573), build (0.572), highlight (0.572), conditional (0.571), oar3 (0.570), coincide (0.568), promising (0.557), harbor (0.556), melanoma (0.555)
STUDY	work (0.674), research (0.646), paper (0.599), integrate (0.579), powerful (0.562), tool (0.556), connect (0.555), explore (0.551), gwa (0.549), valuable (0.548), poorly (0.538), complementary (0.536), current (0.535), far (0.532), focus (0.528), comprehensive (0.525), unravel (0.523), lack (0.522), race (0.521), consistently (0.520)
ASSOCIATE	notably (0.576), significantly (0.548), relate (0.540), enrich (0.512), nearby (0.512), coincide (0.496), conclusion (0.492), affect (0.487), correlate (0.485), contain (0.484), identify (0.479), specifically (0.479), responsible (0.478), summary (0.477), associated (0.476), report (0.474), stature (0.467), centromeric (0.466), vicinity (0.464), locate (0.460)
CHROMOSOME	centromeric (0.720), middle (0.716), mbp (0.712), localize (0.711), harbour (0.706), bta11 (0.689), oar3 (0.688), chr (0.684), distal (0.683), vicinity (0.674), sus (0.672), autosomal (0.658), bta20 (0.653), bta26 (0.653), bta2 (0.649), coincide (0.649), bta6 (0.645), chromosomal (0.644), bta7 (0.643), scrofa (0.641)
ANALYSIS	software (0.694), methodology (0.630), gemma (0.628), multitrait (0.626), comparison (0.600), emmax (0.600), multi (0.598), package (0.597), approach (0.595), conditional (0.595), method (0.595), principal (0.593), combination (0.592), ldla (0.590), chi (0.583), complete (0.579), regression (0.572), concordant (0.572), simple (0.571), multivariate (0.571)

Table 2: Word2Vec baseline (no phrase mining). Each row shows the target word with its top 20 nearest neighbors.

Query Word	Top 20 Most Similar Words (similarity score)
QTL	arm (0.675), qtls (0.662), chromosome (0.650), gga5 (0.636), gga3 (0.633), ssc1 (0.621), ssc3 (0.619), localize (0.617), coincide (0.610), chromosomal (0.609), ssc9 (0.608), gga4 (0.604), imprinting (0.598), localization (0.597), ssc7 (0.597), cm. (0.597), ssc10 (0.597), fat1 (0.591), overlap (0.589), epistatic (0.587)
SNP	bonferroni (0.644), genbank (0.634), intronic (0.628), adjacent (0.624), infer (0.616), thirty (0.614), correction (0.613), 129 (0.610), snv (0.610), untranslated (0.605), nucleotide (0.603), nearby (0.596), glm (0.593), intron (0.587), conditional (0.585), 0.5 (0.585), bovinesnp50 (0.581), flanking (0.575), uncover (0.574), bioinformatics (0.573)
GENE	plausible (0.715), positional (0.686), functional (0.686), gnas (0.661), proximity (0.656), involvement (0.656), regulatory (0.650), upstream (0.649), associated (0.648), element (0.647), annotation (0.647), mitf (0.645), mirna (0.644), list (0.642), literature (0.640), ontology (0.637), function (0.636), spp1 (0.636), directly (0.632), regulate (0.632)
REGION	segment (0.683), vicinity (0.679), mbp (0.675), harbor (0.660), proximal (0.660), encompass (0.659), proximity (0.648), middle (0.641), centromeric (0.638), harbour (0.636), sixteen (0.631), locate (0.618), gnas (0.618), nearby (0.614), bta20 (0.609), overlap (0.608), annotate (0.603), localize (0.594), consistently (0.594), flanking (0.591)
ASSOCIATION	conditional (0.641), gemma (0.599), gwa (0.596), bioinformatics (0.573), genbank (0.560), genomewide (0.557), snp (0.555), chi (0.551), weighted (0.551), gwas (0.551), snv (0.546), nearby (0.545), initial (0.539), bonferroni (0.536), statistically (0.536), consistently (0.533), series (0.524), package (0.523), meta (0.521), glm (0.521)
IDENTIFY	detect (0.674), nearby (0.608), discover (0.608), find (0.607), consistently (0.600), verify (0.593), vicinity (0.588), oar3 (0.585), locate (0.584), centromeric (0.583), notably (0.582), mbp (0.578), build (0.575), uncover (0.574), report (0.572), reveal (0.570), sixteen (0.570), kit (0.568), annotate (0.566), identification (0.564)
STUDY	work (0.663), research (0.617), paper (0.571), powerful (0.570), current (0.567), integrate (0.560), race (0.559), poorly (0.558), recent (0.557), explore (0.553), tool (0.552), connect (0.551), valuable (0.550), obesity (0.548), complementary (0.547), feature (0.547), validate (0.543), unravel (0.543), helpful (0.540), foundation (0.538)
ASSOCIATE	notably (0.591), affect (0.564), significantly (0.538), correlate (0.530), relate (0.525), specifically (0.520), strongly (0.515), coincide (0.512), nearby (0.509), identify (0.497), furthermore (0.497), locate (0.495), influence (0.494), enrich (0.493), interestingly (0.490), contain (0.481), responsible (0.481), bta26 (0.477), wool (0.477), 0.001 (0.476)
CHROMOSOME	harbour (0.742), mbp (0.734), localize (0.734), centromeric (0.719), middle (0.715), bta11 (0.704), gallus (0.695), oar3 (0.687), chr (0.687), bta26 (0.682), vicinity (0.680), bta20 (0.675), autosomal (0.674), sixteen (0.668), distal (0.667), scrofa (0.667), ssc3 (0.667), chromosomal (0.664), autosome (0.664), coincide (0.662)
ANALYSIS	software (0.707), gemma (0.664), comparison (0.638), emmax (0.620), multitrait (0.615), methodology (0.613), chi (0.606), concordant (0.605), multi (0.601), package (0.597), conditional (0.596), univariate (0.593), principal (0.592), method (0.590), separately (0.588), ldla (0.588), approach (0.588), ibd (0.583), powerful (0.580), identity (0.579)

Table 3: Word2Vec with Phrase Mining Method 1. Each row shows the target word with its top 20 nearest neighbors.

Query Word	Top 20 Most Similar Words (similarity score)
QTL	qtl_trait (0.663), qtl_chromosome (0.658), additional_qtl (0.647), qtl_position (0.637), qtl_find (0.634), significant_chromosome_wide_level (0.631), result_confirm (0.629), previously_report_qtl (0.628), trait_qtl (0.626), qtl_affect_trait (0.624), fat_trait (0.624), qtl_detect (0.624), arm (0.624), highly_significant_qtl (0.623), chromosomal (0.622), pleiotropic_effect (0.620), genome_wise_significant_qtl (0.620), chromosome_qtl (0.616), qtl_identify (0.614), qtls (0.613)
GENE	igf1 (0.779), transcription_factor_bind_site (0.764), gene_snp (0.763), upstream (0.759), nr6a1 (0.757), gene_include (0.756), nucleotide (0.750), el_ment (0.750), mirna (0.748), associate_milk_production (0.745), positional_candidate (0.744), gnas (0.743), mitf (0.740), positional_functional_candidate_gene (0.739), abcg2 (0.735), gene_find (0.729), amino_acid_substitution (0.727), transcription_factor (0.727), gene_identify (0.725), select_candidate_gene (0.725)
SNP	snp_marker (0.728), snp_identify (0.720), snp_locate (0.716), haplotype_block (0.714), find_significantly_associate (0.706), adjacent (0.702), total_snp (0.700), snp_snp (0.698), analysis_snp (0.694), snp_find (0.685), bonferroni_correction (0.685), glm (0.681), snp_chromosome (0.680), slide (0.676), window (0.673), snp_detect (0.669), statistical_analysis (0.668), association_analysis_indicate (0.665), reveal_significant_association (0.664), association_snp (0.657)
TRAIT	affect_trait (0.586), phenotypic (0.563), trait_include (0.561), distinguish (0.558), genetic_correlation (0.556), multiple (0.556), production_trait (0.556), shape (0.555), example (0.554), large_number (0.552), qtl_trait (0.552), loci_affect (0.549), parameter (0.548), parasite_resistance (0.546), ease (0.544), identify_qtl (0.544), gestation_length (0.540), qtl_find (0.539), carcass_composition_trait (0.538), fraction (0.536)
GENOTYPE	animal_genotype (0.692), individual_genotype (0.688), snp_genotype (0.685), mutant (0.642), minor_allele_frequency (0.642), china (0.639), impute (0.638), genotypes (0.637), result_show (0.629), significantly_high (0.624), genotype_snp (0.623), 200 (0.623), trait_analyze (0.621), ovinesnp50 (0.617), amplify (0.617), affymetrix (0.616), polymorphic (0.612), density_single_nucleotide_polymorphism (0.610), statistical_analysis (0.608), p<0.05 (0.604)
PIG	swine (0.704), iberian (0.701), pig_population (0.700), duroc_pig (0.699), gilt (0.673), sutai (0.664), commercial_pig (0.657), meishan (0.651), f41 (0.650), duroc (0.645), yorkshire (0.642), trait_relate (0.637), large_white_pig (0.635), pietrain (0.634), reproductive_trait (0.629), berkshire (0.629), pig_breed (0.626), teat_number (0.624), genome_wide_analysis (0.622), boar (0.621)
POPULATION	experimental (0.741), outbred (0.725), population_result (0.715), commercial_population (0.715), crossbred (0.700), diverse (0.693), qtl_segregate (0.692), yorkshire (0.682), berkshire (0.677), fine_map_qtl (0.675), scale (0.671), backcross (0.664), originate (0.661), marker_genotype (0.658), synthetic (0.653), jointly (0.653), growth_fatness (0.652), founder (0.651), studied (0.650), sutai (0.650)
BREED	polish (0.694), cattle_breed (0.680), qinchuan (0.669), pig_breed (0.668), luxi (0.666), finnish (0.666), bos (0.663), simmental (0.662), native (0.659), bos_taurus (0.657), jiaxian (0.653), zebu (0.649), holstein_population (0.647), taurine (0.647), indicus (0.646), limousin (0.646), ayrshire (0.641), china (0.639), majority (0.630), asian (0.628)
REGION	region_identify (0.753), proximal (0.717), middle (0.715), flanking (0.711), centromeric (0.709), haplotype_block (0.693), encompass (0.691), mbp (0.684), untranslated (0.679), flank (0.676), segment (0.671), respectively_snp_locate (0.671), region_contain (0.668), genomic_region_identify (0.666), vicinity (0.664), narrow (0.662), quantitative_trait_locus (0.662), downstream (0.661), overlap (0.660), region_bovine_chromosome (0.660)
MARKER	distance (0.776), add (0.671), confidence_interval_qtl (0.656), marker_interval (0.642), marker_genotype (0.641), marker_chromosome (0.639), microsatellite (0.628), linkage_disequilibrium (0.628), span (0.627), cm. (0.625), recombination (0.623), linkage_map (0.621), resolution (0.621), surround (0.618), informative (0.618), initially (0.612), additional (0.612), refined (0.611), marker_association (0.609), bta5 (0.605)

Table 4: Word2Vec with Phrase Mining Method 2. Each row shows the target word with its top 20 nearest neighbors.