Rough Analysis of v2_marked_and_linked_test.json

General Information

- Num. Entries: 7
- Tokenizer used: dmis-lab/biobert-v1.1

Definitions:

- Words represent the number of entities that have been considered distinct during the proccess of assigning tags to part of the text, it is essentially the text split on spaces
- Tokens are the actual pieces of data going into the model, punctuation is split from words and some large words are split down into smaller parts
- Total tags is the total number of tags across all words (If two or more words are part of the same tag this counts as a tag for each word)
- Unique tags is the total number of tags that are unique (not part of same tag)
- Max tags is the most tags a single word has
- Tagged words is the number of words that have at least one tag
- Tagged words % is the percentage of words that are tagged, it's given in brackets for individual entries.
- Avg tags is the average number of tags per word that is tagged
- MC words is the number of words with tags across multiple categories
- Total links is the total number of links across all tags (If two or more tags are linked this counts as a link for each one)
- Unique links is the total number of links that are unique (not part of the same link)
- Max links is the most amount of links a single tag or word has associated with it
- Linked tags, words is the number of tags or words that have links
- Linked % tags, words is the percentage of tags or words with links
- Avg links per tag, word is the average number of links per tag or word
- The schema used is BIES as it is all encompassing and models can convert to a lower resolution if wanted (Beginning, Inner, End, Singleton)

Sizes

- Average words: 219.9
- Maximum words: 250
- Minimum words: 173

- Average tokens: 412.4
- Maximum tokens: 515
- Minimum tokens: 312

- Entries with over 512 tokens: 1/7, 14.29%

Tags

Maximums

- Total tags: 51
- Unique tags: 27

- Max tags: 2
- Tagged words: 50
- Tagged words %: 23.26%
- Avg tags: 1.02
- MC words: 1

Averages

- Total tags: 28.0
- Unique tags: 14.57
- Max tags: 1.14
- Tagged words: 27.86
- Tagged words %: 12.46%
- Avg tags: 1.0
- MC words: 0.14

Links

Maximums

- Total links: 70
- Unique links: 9
- Max links per tag, word: 6, 6
- Linked tags, words: 40, 40
- Linked % tags, words: 100.0%, 17.09%
- Avg links per tag, word: 2.8, 2.8

Averages

- Total links: 26.43
- Unique links: 3.14
- Max links per tag, word: 2.0, 2.0
- Linked tags, words: 17.57, 17.57
- Linked % tags, words: 66.3%, 7.79%
- Avg links per tag, word: 1.33, 1.33

Schema

- Maximums (B, I, E, S): 11, 19, 11, 18
- Averages (B, I, E, S): 6.0, 7.43, 6.0, 8.57

Labels

- Label: Total, Maximum per entry, Average per entry

Categories

- Perturbing_Action: 76, 18, 10.86
- Context: 57, 19, 8.14
- Effect: 28, 9, 4.0
- Phenotype: 35, 11, 5.0

Perturbing_Action

- Gene Loss-Of-Function: 28, 10, 4.0
- Gene Gain-Of-Function: 16, 11, 2.29
- Rnai/Knockdown: 21, 11, 3.0
- Pharmacological Inhibition: 11, 6, 1.57
- Pharmacological Augmentation: 0, 0, 0.0

- Other: 0, 0, 0.0

Context

- Patient: 0, 0, 0.0
- Organism: 8, 5, 1.14
- Tissue/Organ: 5, 3, 0.71
- Neoplasm: 0, 0, 0.0
- Graft: 0, 0, 0.0
- Xenograft: 0, 0, 0.0
- Cells: 10, 6, 1.43
- Transformed Cells: 24, 10, 3.43
- Organoid: 0, 0, 0.0
- In Vitro: 4, 2, 0.57
- In Vivo: 6, 2, 0.86

Effect

- Positive: 13, 5, 1.86
- Negative: 14, 4, 2.0
- Regulates: 1, 1, 0.14
- Rescues: 0, 0, 0.0
- No Effect: 0, 0, 0.0

Phenotype

- Adhesion: 0, 0, 0.0
- Apoptosis: 1, 1, 0.14
- Anoikis: 0, 0, 0.0
- Autophagy: 11, 5, 1.57
- Cell Cycle Arrest: 0, 0, 0.0
- Cell Death: 0, 0, 0.0
- Cell Growth: 0, 0, 0.0
- Cell Survival: 1, 1, 0.14
- Colony Formation: 0, 0, 0.0
- Differentiation: 0, 0, 0.0
- Entosis: 0, 0, 0.0
- Epithelial-Mesenchymal Transition: 0, 0, 0.0
- Ferroptosis: 0, 0, 0.0
- Invasion: 2, 2, 0.29
- Metastasis: 5, 3, 0.71
- Migration: 2, 2, 0.29
- Mitophagy: 0, 0, 0.0
- Necroptosis: 0, 0, 0.0
- Necrosis: 0, 0, 0.0
- Oncosis: 0, 0, 0.0
- Proliferation: 4, 2, 0.57
- Pyroptosis: 0, 0, 0.0
- Quiescence: 0, 0, 0.0
- Self-Renewal: 3, 3, 0.43
- Senescence: 0, 0, 0.0
- Transformation: 0, 0, 0.0
- Tumour Growth: 0, 0, 0.0
- Tumourigenesis: 2, 2, 0.29
- Tumour Initiation: 4, 4, 0.57
- Tumour Progression: 0, 0, 0.0
- Tumour Regression: 0, 0, 0.0

Information on each entry

Title: 33_PMID31223056.txt_CC2.xml, Words: 250, Tokens: 515
Tag data: Total: 24, Unique: 15, Max tags: 1, Tagged words: 24 (10%), Avg tags: 1.0, MC words: 0
Link data: Total: 5, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (5 (20.83%), 5 (2.0%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (4, 5, 4, 11)

Title: 102_PMID32054768.txt_CC2.xml, Words: 234, Tokens: 429
Tag data: Total: 46, Unique: 16, Max tags: 1, Tagged words: 46 (20%), Avg tags: 1.0, MC words: 0
Link data: Total: 54, Unique: 4, Max links (tag, word): (2, 2), Linked (tags, words): (40 (86.96%), 40 (17.09%))
Avg links (tag, word): (1.35, 1.35), Schema counts (B, I, E, S): (11, 19, 11, 5)

Title: 112_PMID33051252.txt_CC2.xml, Words: 238, Tokens: 414
Tag data: Total: 25, Unique: 15, Max tags: 1, Tagged words: 25 (11%), Avg tags: 1.0, MC words: 0
Link data: Total: 70, Unique: 9, Max links (tag, word): (6, 6), Linked (tags, words): (25 (100.0%), 25 (10.5%))
Avg links (tag, word): (2.8, 2.8), Schema counts (B, I, E, S): (6, 4, 6, 9)

Title: 64_PMID31775562.txt_CC2.xml, Words: 215, Tokens: 424
Tag data: Total: 51, Unique: 27, Max tags: 2, Tagged words: 50 (23%), Avg tags: 1.02, MC words: 1
Link data: Total: 19, Unique: 2, Max links (tag, word): (1, 1), Linked (tags, words): (19 (37.25%), 19 (8.84%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (9, 15, 9, 18)

Title: 54_PMID31373534.txt_CC2.xml, Words: 233, Tokens: 448
Tag data: Total: 29, Unique: 18, Max tags: 1, Tagged words: 29 (12%), Avg tags: 1.0, MC words: 0
Link data: Total: 24, Unique: 4, Max links (tag, word): (2, 2), Linked (tags, words): (21 (72.41%), 21 (9.01%))
Avg links (tag, word): (1.14, 1.14), Schema counts (B, I, E, S): (5, 6, 5, 13)

Title: 231_PMID30692207.txt_CC2.xml, Words: 173, Tokens: 345
Tag data: Total: 15, Unique: 7, Max tags: 1, Tagged words: 15 (9%), Avg tags: 1.0, MC words: 0
Link data: Total: 7, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (7 (46.67%), 7 (4.05%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (5, 3, 5, 2)

Title: 194_PMID32699135.txt_CC2.xml, Words: 196, Tokens: 312
Tag data: Total: 6, Unique: 4, Max tags: 1, Tagged words: 6 (3%), Avg tags: 1.0, MC words: 0
Link data: Total: 6, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (6 (100.0%), 6 (3.06%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (2, 0, 2, 2)