

Rough Analysis of v2_all_marked_test.json

General Information

- Num. Entries: 24
- Tokenizer used: dmis-lab/biobert-v1.1

Definitions:

- Words represent the number of entities that have been considered distinct during the process 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: 209.8
- Maximum words: 269
- Minimum words: 151
- Average tokens: 399.4
- Maximum tokens: 536
- Minimum tokens: 271
- Entries with over 512 tokens: 4/24, 16.67%

Tags

Maximums

- Total tags: 55
- Unique tags: 27

- Max tags: 2
- Tagged words: 53
- Tagged words %: 20.7%
- Avg tags: 1.04
- MC words: 0

Averages

- Total tags: 18.21
- Unique tags: 9.46
- Max tags: 1.0
- Tagged words: 18.12
- Tagged words %: 8.26%
- Avg tags: 0.96
- MC words: 0.0

Links

Maximums

- Total links: 54
- Unique links: 4
- Max links per tag, word: 2, 2
- Linked tags, words: 40, 40
- Linked % tags, words: 91.67%, 17.09%
- Avg links per tag, word: 1.57, 1.57

Averages

- Total links: 9.42
- Unique links: 1.08
- Max links per tag, word: 0.75, 0.75
- Linked tags, words: 7.54, 7.54
- Linked % tags, words: 27.17%, 3.23%
- Avg links per tag, word: 0.58, 0.58

Schema

- Maximums (B, I, E, S): 11, 21, 11, 20
- Averages (B, I, E, S): 3.92, 4.83, 3.92, 5.54

Labels

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

Categories

- Perturbing_Action: 185, 21, 7.71
- Context: 137, 19, 5.71
- Effect: 49, 7, 2.04
- Phenotype: 66, 9, 2.75

Perturbing_Action

- Gene Loss-Of-Function: 79, 15, 3.29
- Gene Gain-Of-Function: 27, 11, 1.12
- Rnai/Knockdown: 48, 13, 2.0
- Pharmacological Inhibition: 18, 6, 0.75
- Pharmacological Augmentation: 4, 4, 0.17

- Other: 9, 8, 0.38

Context

- Patient: 0, 0, 0.0
- Organism: 14, 5, 0.58
- Tissue/Organ: 9, 3, 0.38
- Neoplasm: 7, 3, 0.29
- Graft: 0, 0, 0.0
- Xenograft: 0, 0, 0.0
- Cells: 54, 15, 2.25
- Transformed Cells: 29, 10, 1.21
- Organoid: 0, 0, 0.0
- In Vitro: 12, 4, 0.5
- In Vivo: 12, 2, 0.5

Effect

- Positive: 17, 4, 0.71
- Negative: 26, 4, 1.08
- Regulates: 0, 0, 0.0
- Rescues: 5, 4, 0.21
- No Effect: 1, 1, 0.04

Phenotype

- Adhesion: 0, 0, 0.0
- Apoptosis: 1, 1, 0.04
- Anoikis: 0, 0, 0.0
- Autophagy: 26, 5, 1.08
- Cell Cycle Arrest: 0, 0, 0.0
- Cell Death: 4, 2, 0.17
- Cell Growth: 2, 2, 0.08
- Cell Survival: 1, 1, 0.04
- Colony Formation: 0, 0, 0.0
- Differentiation: 1, 1, 0.04
- Entosis: 0, 0, 0.0
- Epithelial-Mesenchymal Transition: 0, 0, 0.0
- Ferroptosis: 0, 0, 0.0
- Invasion: 0, 0, 0.0
- Metastasis: 5, 3, 0.21
- Migration: 0, 0, 0.0
- Mitophagy: 2, 2, 0.08
- Necroptosis: 0, 0, 0.0
- Necrosis: 0, 0, 0.0
- Oncosis: 0, 0, 0.0
- Proliferation: 2, 2, 0.08
- Pyroptosis: 0, 0, 0.0
- Quiescence: 0, 0, 0.0
- Self-Renewal: 0, 0, 0.0
- Senescence: 6, 4, 0.25
- Transformation: 0, 0, 0.0
- Tumour Growth: 7, 2, 0.29
- Tumourigenesis: 7, 2, 0.29
- Tumour Initiation: 2, 2, 0.08
- Tumour Progression: 0, 0, 0.0
- Tumour Regression: 0, 0, 0.0

Information on each entry

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: 38_PMID31679456.txt_CC2.xml, Words: 256, Tokens: 486

Tag data: Total: 55, Unique: 27, Max tags: 2, Tagged words: 53 (21%), Avg tags: 1.04, MC words: 0

Link data: Total: 33, Unique: 3, Max links (tag, word): (2, 2), Linked (tags, words): (21 (38.18%), 21 (8.2%))

Avg links (tag, word): (1.57, 1.57), Schema counts (B, I, E, S): (7, 21, 7, 20)

Title: 206_PMID32883681.txt_CC2.xml, Words: 151, Tokens: 319

Tag data: Total: 4, Unique: 4, Max tags: 1, Tagged words: 4 (3%), Avg tags: 1.0, MC words: 0

Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))

Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (0, 0, 0, 4)

Title: 23_PMID30208757.txt_CC2.xml, Words: 170, Tokens: 320

Tag data: Total: 6, Unique: 2, Max tags: 1, Tagged words: 6 (4%), Avg tags: 1.0, MC words: 0

Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))

Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (2, 2, 2, 0)

Title: 209_PMID30692202.txt_CC2.xml, Words: 173, Tokens: 308

Tag data: Total: 5, Unique: 1, Max tags: 1, Tagged words: 5 (3%), Avg tags: 1.0, MC words: 0

Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))

Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (1, 3, 1, 0)

Title: 30_PMID30681394.txt_CC2.xml, Words: 163, Tokens: 352

Tag data: Total: 5, Unique: 4, Max tags: 1, Tagged words: 5 (3%), Avg tags: 1.0, MC words: 0

Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))

Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (1, 0, 1, 3)

Title: 18_PMID31931659.txt_CC2.xml, Words: 247, Tokens: 534

Tag data: Total: 40, Unique: 24, Max tags: 1, Tagged words: 40 (16%), Avg tags: 1.0, MC words: 0

Link data: Total: 26, Unique: 3, Max links (tag, word): (2, 2), Linked (tags, words): (19 (47.5%), 19 (7.69%))

Avg links (tag, word): (1.37, 1.37), Schema counts (B, I, E, S): (6, 10, 6, 18)

Title: 261_PMID30692209.txt_CC2.xml, Words: 166, Tokens: 271

Tag data: Total: 22, Unique: 11, Max tags: 1, Tagged words: 22 (13%), Avg tags: 1.0, MC words: 0

Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))

Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (6, 5, 6, 5)

Title: 70_PMID31448672.txt_CC2.xml, Words: 179, Tokens: 365

Tag data: Total: 22, Unique: 10, Max tags: 1, Tagged words: 22 (12%), Avg tags: 1.0, MC words: 0

Link data: Total: 7, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (7 (31.82%), 7 (3.91%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (6, 6, 6, 4)

Title: 162_PMID32737118.txt_CC2.xml, Words: 236, Tokens: 412

Tag data: Total: 16, Unique: 11, Max tags: 1, Tagged words: 16 (7%), Avg tags: 1.0, MC words: 0

Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))

Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (5, 0, 5, 6)

Title: 15_PMID32249716.txt_CC2.xml, Words: 240, Tokens: 429

Tag data: Total: 26, Unique: 17, Max tags: 1, Tagged words: 26 (11%), Avg tags: 1.0, MC words: 0

Link data: Total: 20, Unique: 3, Max links (tag, word): (2, 2), Linked (tags, words): (18 (69.23%), 18 (7.5%))

Avg links (tag, word): (1.11, 1.11), Schema counts (B, I, E, S): (5, 4, 5, 12)

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: 45_PMDID30686098.txt_CC2.xml, Words: 238, Tokens: 534
Tag data: Total: 11, Unique: 6, Max tags: 1, Tagged words: 11 (5%), Avg tags: 1.0, MC words: 0
Link data: Total: 7, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (7 (63.64%), 7 (2.94%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (4, 1, 4, 2)

Title: 248_PMDID32217667.txt_CC2.xml, Words: 199, Tokens: 324
Tag data: Total: 0, Unique: 0, Max tags: 0, Tagged words: 0 (0%), Avg tags: 0, MC words: 0
Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0%), 0 (0%))
Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (0, 0, 0, 0)

Title: 293_PMDID31123067.txt_CC2.xml, Words: 173, Tokens: 293
Tag data: Total: 5, Unique: 2, Max tags: 1, Tagged words: 5 (3%), Avg tags: 1.0, MC words: 0
Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))
Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (1, 2, 1, 1)

Title: 42_PMDID31276435.txt_CC2.xml, Words: 184, Tokens: 368
Tag data: Total: 12, Unique: 7, Max tags: 1, Tagged words: 12 (7%), Avg tags: 1.0, MC words: 0
Link data: Total: 5, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (5 (41.67%), 5 (2.72%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (4, 1, 4, 3)

Title: 217_PMDID31857346.txt_CC2.xml, Words: 166, Tokens: 316
Tag data: Total: 9, Unique: 4, Max tags: 1, Tagged words: 9 (5%), Avg tags: 1.0, MC words: 0
Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))
Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (3, 2, 3, 1)

Title: 196_PMDID32213544.txt_CC2.xml, Words: 196, Tokens: 392
Tag data: Total: 15, Unique: 8, Max tags: 1, Tagged words: 15 (8%), Avg tags: 1.0, MC words: 0
Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))
Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (4, 3, 4, 4)

Title: 102_PMDID32054768.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: 141_PMDID32046982.txt_CC2.xml, Words: 238, Tokens: 360
Tag data: Total: 12, Unique: 4, Max tags: 1, Tagged words: 12 (5%), Avg tags: 1.0, MC words: 0
Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))
Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (4, 4, 4, 0)

Title: 166_PMDID32060146.txt_CC2.xml, Words: 269, Tokens: 431
Tag data: Total: 24, Unique: 10, Max tags: 1, Tagged words: 24 (9%), Avg tags: 1.0, MC words: 0
Link data: Total: 29, Unique: 3, Max links (tag, word): (2, 2), Linked (tags, words): (22 (91.67%), 22 (8.18%))
Avg links (tag, word): (1.32, 1.32), Schema counts (B, I, E, S): (7, 7, 7, 3)

Title: 5_PMDID30870073.txt_CC2.xml, Words: 242, Tokens: 491
Tag data: Total: 14, Unique: 6, Max tags: 1, Tagged words: 14 (6%), Avg tags: 1.0, MC words: 0
Link data: Total: 5, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (5 (35.71%), 5 (2.07%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (3, 5, 3, 3)

Title: 289_PMDID30692206.txt_CC2.xml, Words: 174, Tokens: 353
Tag data: Total: 21, Unique: 9, Max tags: 1, Tagged words: 21 (12%), Avg tags: 1.0, MC words: 0
Link data: Total: 11, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (11 (52.38%), 11 (6.32%))
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (4, 8, 4, 5)

Title: 68_PMDID32186433.txt_CC2.xml, Words: 259, Tokens: 536
Tag data: Total: 14, Unique: 11, Max tags: 1, Tagged words: 14 (5%), Avg tags: 1.0, MC words: 0
Link data: Total: 0, Unique: 0, Max links (tag, word): (0, 0), Linked (tags, words): (0 (0.0%), 0 (0%))
Avg links (tag, word): (0, 0), Schema counts (B, I, E, S): (1, 2, 1, 10)