話說天下大勢分久必合合久必分

Word-based tokenization

話説 天下 大 勢 分 久 必 合

合 久 必 分

Overlapping 3–gram tokenization

話説天　下大勢　分久必　合合久

説天下　大勢分　久必合　合久必

天下大　勢分久　必合合　久必分

1-gram tokenization

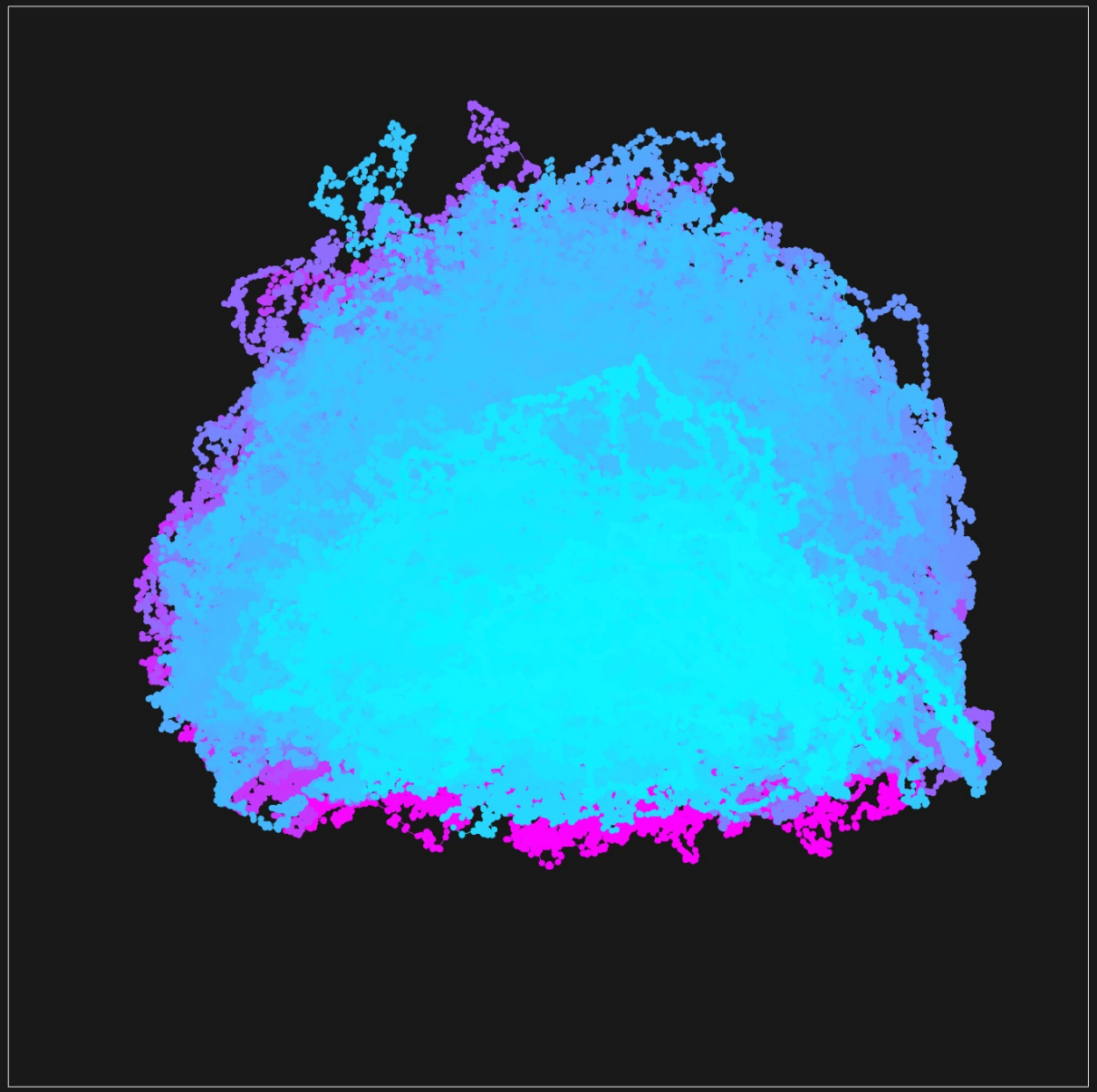
話 説 天 下 大 勢 分 久 必 合

合 久 必 分

Sub-character bytecode tokenization

話	說	天	下
e8a9b1	e8aaaa	e5a4a9	e4b88b

大	勢	分
e5a4a7	e58ba2	e58886



500字 Sliding window

1000 Most Frequent 字

1st dot: 1-501
2nd dot: 2-502
...
Last dot: 751348-751848