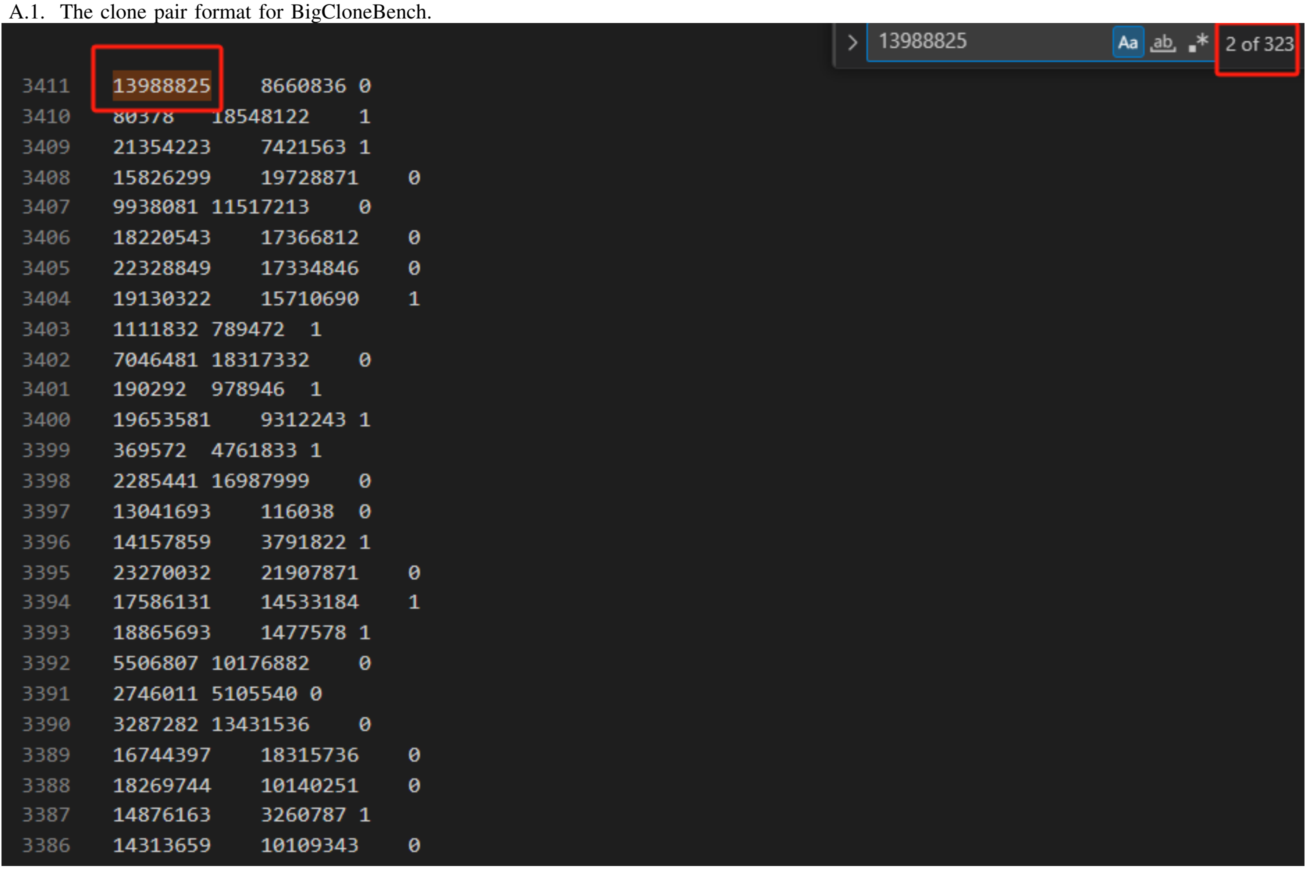
Appendix: TransformCode: A Contrastive Learning Framework for Code Embedding via Subtree Transformation

Zixiang Xian, Rubing Huang, Dave Towey, Chunrong Fang, Zhenyu Chen

Figure A.1 shows how the pairs were formatted in the BigCloneBench dataset for unsupervised code-clone detection. We used 7302 unique training samples from the BigCloneBench dataset, which were formatted in pairs to include both clones and non-clones. Our TransformCode framework was trained on the unique samples, and the same dataset settings were applied to other unsupervised tasks, maintaining consistency with the BigCloneBench dataset.

Fig. A.1. The clone pair format for BigCloneBench.



```
13988825	8660836	0
80378	18548122	1
21354223	7421563	1
15826299	19728871	0
9938081	11517213	0
18220543	17366812	0
22328849	17334846	0
19130322	15710690	1
1111832	789472	1
7046481	18317332	0
190292	978946	1
19653581	9312243	1
369572	4761833	1
2285441	16987999	0
13041693	116038	0
14157859	3791822	1
23270032	21907871	0
17586131	14533184	1
18865693	1477578	1
5506807	10176882	0
2746011	5105540	0
3287282	13431536	0
16744397	18315736	0
18269744	10140251	0
14876163	3260787	1
14313659	10109343	0
```