ما هي البيانات الضخمة "البيغ داتا" (Big Data)؟



نعيش الآن في عصر المعلومات، ومعظم ما نقوم به يتأثر بشكلٍ كبير بقدرتنا على الوصول إلى كميات هائلة من البيانات سواء أكان ذلك عبر الإنترنت، أم حواسيبنا، أم هواتفنا المحمولة. والكلمة الطنانة التي تصف هذا الكمّ من المعلومات هي البيانات الضخمة (**Big Data**).

وفي العام 2012 حددت الحكومة البريطانية البيانات الضخمة بوصفها واحدة من ثمان تقنيات مستقبلية عظيمة، ولذلك فما هو التحدي الذي تطرحه البيانات الضخمة؟ وكيف يُمكننا مواجهته؟

من أين أتت البيانات الضخمة؟

ربما يكون المصدر الرئيسي للبيانات الضخمة حالياً هو الإنترنت، ووفقاً لتقدير حديث، فإنّ حوالي 21^10 بايت (أي زيتا بايت **zettabyte**) من المعلومات تُضاف إلى الإنترنت كل عام، ويكون معظمها على شكل محتوى غرافيكي. وتتجاوز تغطية الإنترنت في المملكة المتحدة 80%، ولكنها لا يتجاوز 20% في بعض الدول.

توضّح هذه الخريطة العدد الإجمالي لمستخدمي الإنترنت في الدولة، إضافة إلى نسبة التغطية بالنسبة للتعداد السكاني الذي يصل إليه الإنترنت في العام 2011.

إنّ المصدر الرئيسي لهذه البيانات، الذي يستمر في النمو، هو المحتوى الموجود على مواقع التواصل الاجتماعي. فعلى سبيل المثال لدى موقع فيسبوك (facebook) الذي انطلق في العام 2004 حوالي 2 مليار مستخدم مسجّل (ما يُعادل ربع عدد سكان العالم)، منهم 1.5 مليار مستخدم نشط. ويوميًا يُضاف إلى هذا الموقع حوالي 2.5 مليار محتوى (أي ما يُعادل 500 تيرابايت من المعلومات)، معظمها يُخزّن على شكل صور. ويُقدر أنّ محرك البحث غوغل يُجري عمليات البحث عن المعلومات الموجودة في 15 إكسابايت (**exabytes**) ـأي 15^10 بايتـ من البيانات، ويُقوم بتلك المهمة بالاعتماد على خوارزمية رياضية ذكية.

للبيانات الضخمة مصدرٌ آخر وهو الهواتف المحمولة والذكية، فاليوم يتجاوز عدد من الهواتف المحمولة في العالم عدد سكانه، مع احتمالية وجود 25000000000000000000 محادثة في الوقت نفسه. وستقدم الخطة المستقبلية المتعلقة بوجود شبكة 5G قدرة وصول إلى البيانات بمعدل 1 غيغابايت في الثانية الواحدة بالنسبة لعشرات العاملين في الوقت نفسه وفي نفس المكتب. وتظهر الحسّاسات التي يُمكنها تقديم مراقبة مستمرة لحالتنا الصحية ـ مع وجود آثار أخلاقية معتبرةـ كإحدى التقنيات التي تتقدم بسرعة في هذا المجال. وستدعم شبكة 5G بضع مئات الآلاف من عمليات الاتصال المتزامنة لعمليات نشر ضخمة لتلك الحساسات. وفي الواقع، فالمستقبل ليس بعيدا، وقريباً ستتصل أجهزتنا مع بعضها البعض بوجود القليل من التدخّل البشري، أو حتى بدونه، فعلى سبيل المثال ستتحدث آلة الطبخ مع آلة الجلي والتنظيف، ومع السوبرماركت أيضا في كل لحظة يتم فيها تحضير وجبة، وهذا ما يُعرف بإنترنت الأشياء (**Internet of things**).



ساعات ذكية تُعطيك المعلومات الصحية الأحدث الخاصة بك أينما كنت

تأتي كمية معتبرة من البيانات المهمة بالنسبة للعلوم الاجتماعية من الطريقة التي نستخدمها للتعامل مع أجهزتنا، والمعلومات التي تُعطيها تلك الأجهزة عن نمط حياتنا. ففي كل مرة نشتري فيها شيئاً ما من أمازون (**Amazon**)، أو نستخدم حسابنا البنكي، أو نُشغّل جهازا كهربائيا، أوهاتفنا، أو نكتب بريدا إلكترونيا، فإننا نُنشئ بيانات ستحتوي معلومات يُمكن من حيث المبدأ تحليلها. وعلى سبيل المثال يُمكن تحديد عادات الشراء، أو المواقع المتتبعة وتسجيلها. ويُمكن استخدام الرياضيات في كل مراحل هذه العملية، لكن يجب ألا نفقد أبدا البعد الأخلاقي عند القيام بذلك.

طبيعة البيانات الضخمة

كانت البيانات الضخمة محط اهتمام ودراسة الرياضيات للأعوام المئة الأخيرة بشكلٍ من الأشكال. وكمثال كلاسيكي على ذلك نجد علم الأرصاد الجوية (meteorology) الذي نحتاج فيه إلى كميات هائلة من الأعداد التي يجب ضغطها لتوليد تنبؤات طقسية واقعية. وبشكلٍ مشابه تنتج مجموعات البيانات الضخمة عن النماذج المناخية، والجيوفيزياء، وعلم الفلك.

وعلى أية حال، فإنّ مجموعات البيانات الموجودة في هذه المسائل ـ على الرغم من ضخامتها ـ مرتبة بشكلٍ جيد ومفهومة أيضا، مع وجود مستويات معروفة من الارتياب طبعا. وهو أمرٌ ناتج عن كونها قادمة من عمليات فيزيائية يفهمها العلماء جيداً. تكمن التحديات الحقيقية في فهم البيانات الضخمة والتعامل معها في العلوم الإحيائية، والعلوم الاجتماعية، وبشكلٍ خاص تلك المبنية على النشاط

الإنساني. وغالبا ما تكون مثل هذه البيانات مشوّهة، وغير كاملة، ولا يُمكن الاعتماد عليها، ومعقدة، وقصصية بدلاً من كونها الأشياء نفسها. أما البيانات الفيزيائية فليست كذلك.

ما هي الأسئلة التي يجب طرحها حول البيانات الضخمة؟

كيف يُمكن تصوير البيانات الضخمة؟ ووضع التخمينات انطلاقاً منها؟ وكيف نُحاكيها ونفهمها؟ وكيف نُجري التجارب على الأنظمة التي ولّدتها، وفي نهاية المطاف كيف بمقدورنا التحكم بمثل هذه الأنظمة؟ إنّ التحديات العلمية والرياضية الكامنة وراء هذه الأسئلة متغيرة بمقدار ما هي مهمة، كما أنّ حجم البيانات الضخم يجعل من عملية الأتمتة حتمية. وهذه الأتمتة (**automation**) تعتمد على الخوارزميات الرياضية.

تتضمن الأسئلة التي قد نسألها بخصوص البيانات الضخمة ما يلي:

- كيف نُصنّف أهمية المعلومات في شبكات واسعة موجودة في متصفحات الإنترنت مثل غوغل؟
- كيف نُحدد فعاليات المستهلكين، وولاءهم وحتى مشاعرهم، وكيف نُجري عمليات الاقتراح الشخصية؟
- كيف نُحاكي الارتيابات الموجودة في الاتجاهات الصحيّة للمرضى الأفراد؟
- كيف نُنجز ونتعامل مع عملية المراقبة الصحية بالوقت الحقيقي خصوصا في البيئة التي سيقودنا إليها انترنت 5G؟
- كيف نستخدم البيانات الذكية في مزودات الطاقة؟

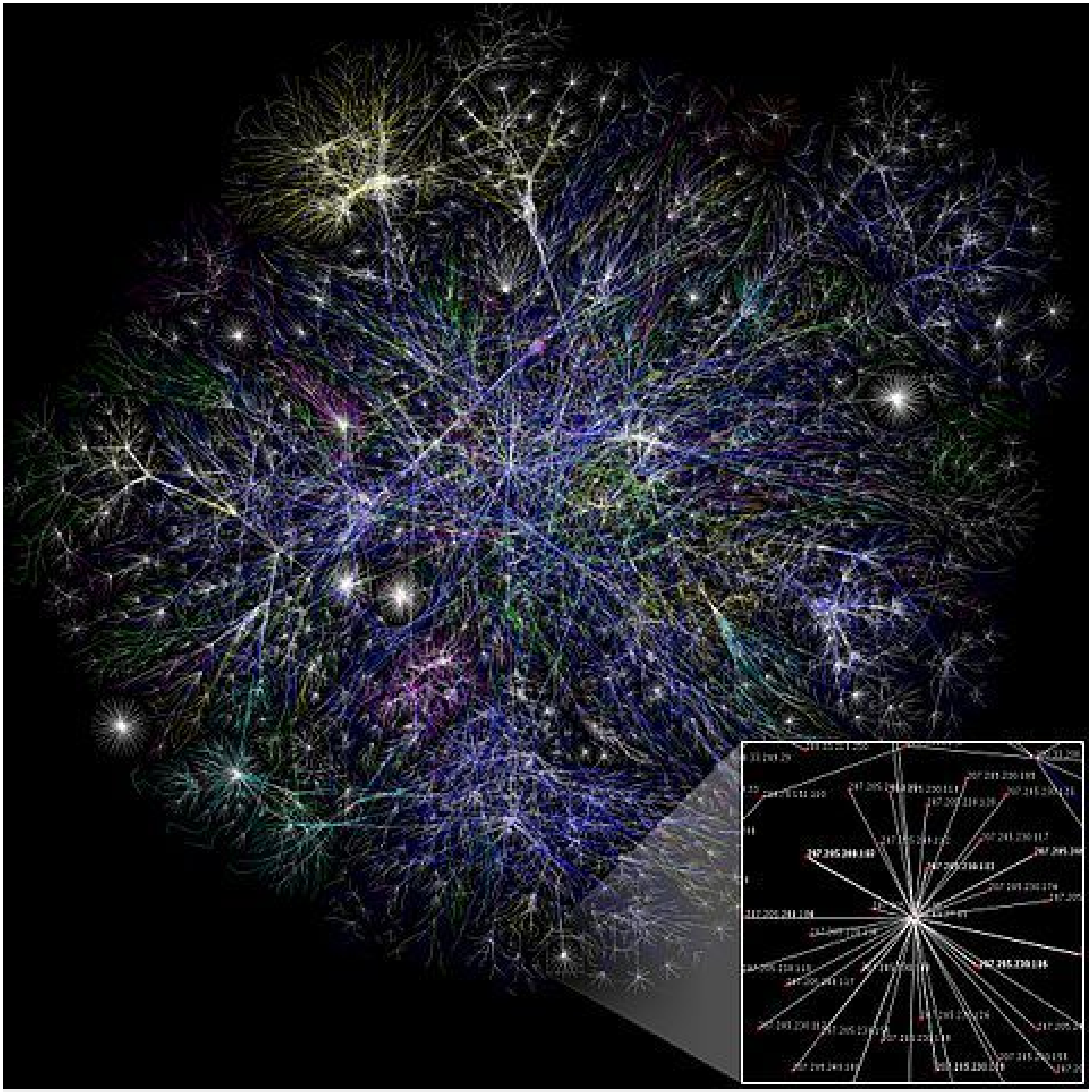
على ما أعتقد فإنّ من العدل القول بأنّ العديد من التطورات المستقبلية في مجال الرياضيات الحديثة (جنباً إلى جنب مع علوم الحاسوب) ستقودنا إما إلى عمليات محاكاة من قبل تطبيقات البيانات الضخمة، أو أنها ستعتمد على الحاجة لفهم البيانات الضخمة. إنّ العديد من التقنيات الرياضية الموجودة حاليا (وبعضها لم يُعتبر رياضيات بحتة حتى وقت حديث) تجد الآن تطبيقاتٍ عملية لها في مجال فهمنا للبيانات الضخمة، وكمثال رئيسي على ذلك نذكر نظرية الشبكات (**network theory**)

شبكات في كل مكان

تصف نظرية الشبكات – كما يقترح الاسم- الأجسام المعروفة بالعُقَد (**nodes**) والمرتبطة مع بعضها البعض عبر ما يُعرف بالحواف (**edges**). يُمكن أن تكون هذه العقد حواسيب، أو شبكات عنكبوتية. أما الحواف فهي الوصلات الموجودة بين الحواسيب، أو الروابط بين مواقع الإنترنت. كما يُمكن أن تكون تلك العقد أيضاً البشر، أما الوصلات فهي في هذه الحالة أصدقاؤهم على الفيسبوك، أو تويتر. أو يُمكن أن تكون مجموعات الهواتف المحمولة، وتكون الروابط في هذه الحالة هي المحادثات أو ببساطة الجوار القريب الذي قد يقود إلى تداخل. تشرح نظرية الشبكات طبيعة الشبكات وتسمح لنا بالبحث عن الوصلات الكائنة بين النقاط المفردة لمجموعات البيانات، ويُمكنها وصف حركة المعلومات حول الشبكة.

في الواقع، تُعتبر عملية إدارة شبكة الهواتف المحمولة (المستخدمة في الواقع في تحميل البيانات أيضاً) مهمة جدا وتطبيقا مستمرا بالنمو لمجال نظرية تلوين الرسوم (**graph colouring**): إيجاد طرق لتلوين الحواف أو العقد الموجودة في الشبكة وفقاً لقيود محددة مثل وجوب امتلاك العقد المتجاورة لألوان مختلفة. فعلى سبيل المثال قد تُمثّل تلك الألوان الترددات المنسوبة لمُرسلات الهواتف المحمولة التي يجب اختيارها بحيث تُقلل من عملية التداخل، ومن ثَمّ يجب أن تكون مختلفة بالنسبة للمرسلات المتجاورة. ولم يتم حتى وقتٍ متأخر النظر إلى تلوين الرسوم على أنه يعود إلى مجال الرياضيات البحتة.

تتضمن الأمثلة الأخرى التي تقود إلى البيانات الضخمة عملية الشبكات التنظيمية مثل شبكات الإدارة، وعصابات الجرائم، وحتى سلوك التصويت في مسابقة الأغنية الأوروبية، إضافة إلى الشبكات التكنولوجية مثل شبكات الطاقة والدارات الكهربائية، وشبكات المعلومات المصنوعة من تفاعلات البروتين-بروتين والجينات ونشر المعلومات والأساطير والشائعات، وأيضاً شبكات النقل مثل الخطوط الجوية، والخدمات اللوجستية الغذائية وأنظمة القطارات تحت الأرض وفوقها، والشبكات البيئية، مثل سلاسل الغذاء والأمراض وآليات العدوى.



خريطة جزئية للإنترنت تعتمد على بيانات العام 2005

قوة نظرية الشبكات

تستطيع هذه النظرية التصدي للعديد من الأسئلة المتعلقة بالبيانات الضخمة، فعندما تتعامل مع شبكات كبيرة جداً لن يكون من السهل دوماً تحديد التكتلات (**clusters**) ـ مجموعات من العقد المرتبطة داخلياً بقوة ـ أو تقسيم البيانات إلى مجموعات تتشارك مميزات شائعة. ومثل هذا النوع من المعلومات مهم جدا وجوهري في مجال التنقيب عن البيانات (**data mining**) والتعرّف على الأنماط (**pattern recognition**). يرتبط هذا الأمر بشكلٍ خاص بقطاع التجزئة الذي يهتم بسلوك وفعاليات المستهلكين، لكنه يُمكن أن يرتبط أيضاً بإيجاد أنماط التصويت في مسابقة الأغنية الأوروبية، وتُقدم نظرية الشبكات الخوارزميات اللازمة لتحديد التكتلات ولتقسيم البيانات.

يساعد مثل هذا التحليل في حل مسألة أخرى مهمة يمكن مواجهتها في العديد من التطبيقات: ربط البيانات التي تعتمد على مستويات مختلفة من التقسيمات في الفضاء والزمن. وكمثال على ذلك نذكر عملية التنبؤ بالطقس، ففيها ربما تأتي بعض البيانات من الأقمار الصناعية التي تدور حول الأرض وتنقل بيانات يصل حجمها إلى عدة ميغا بايت في الثانية الواحدة. وقد ينجم قسم آخر من البيانات عن الأفراد الموجودين في محطات أرضية معزولة الذين يُقدمون بضعة قياسات كل يوم. كما أن بعضاً من تلك البيانات قد يكون تاريخيا مثل سجلات قباطنة البحار التي تشمل الأعوام المئة الأخيرة. ومجموعات البيانات الثلاث هذه مفيدة ويجب ربطها معاً بطريقة سلسلة وواضحة.

تتمتع كيفية وصل الشبكة أو انتشار الوصلات بنفس الأهمية؟ إضافة إلى أقصر المسارات داخل الشبكة.

هذه الأسئلة جوهرية للحصول على إنترنت فعال، إضافة إلى تفسير البيانات اللوجستية، وفهم الاتصالات السريعة، وأيضاً التسويق. تُعتبر نظرية الشبكات أساسية أيضاً في مجال البحث عن العقد المؤثرة في الشبكات العملاقة، فالعقد ذات الاتصالات القوية ـ سواء مثلت الأشخاص أو مواقع الانترنت، أو المطارات ـ هائلة الأهمية بالنسبة لتماسك الشبكة؛ لأنّ حذفها سيؤثر بشكلٍ كبير على الإتاحية الإجمالية للاتصالات. وبالتالي يُمكن استخدام مثل هذا النوع من المعلومات لتحطيم التنظيمات الإرهابية، وإيقاف انتشار الأوبئة، أو الحفاظ على الحركة الجوية عندما تتأثر المنطقة بطقسٍ سيء.



المعلومات والأمراض المعدية والنميمة كلها منتشرة في فضاء الشبكات الاجتماعية

ماذا يُمكن للرياضيات أن تُنجز أيضا؟

تُمثّل نظرية المعلومات إحدى التقنيات الرياضية المستخدمة في دراسة البيانات الضخمة، ويأخذ معظم البيانات الضخمة شكل صور، ولذلك فإنّ الخوارزميات الرياضية التي تُصنِّف وتفسر وتحلل وتضغط الصور غاية في الأهمية. إضافة إلى ذلك فطالما تم استخدام الطرق الإحصائية في تحليل وتفسير الصور، لكنها حصلت حديثاً على دفعة مهمة بفضل خوارزميات رياضية جديدة أصبحت مؤخراً جزءا من الرياضيات البحتة رغم اعتقاد الناس حتى وقتٍ قريب بعدم وجود أي تطبيقات مباشرة لها في العالم الحقيقي.

بعضٌ من تلك الخوارزميات مبني على تحليل المعادلات المعقدة، مما يقود إلى الحصول على تطبيقات قوية وغير متوقعة لأدوات عالية التقنية انطلاقاً من النظريات المتعلقة بتلك المعادلات. تلعب الطبولوجيا الجبرية (**Algebraic topology**) ـ وهي حقل الرياضيات الذي يدرسُ خواص الأشكال باستخدام الجبر ـ دوراً مفيداً جداً في تصنيف الصور. كما أنّ التقنيات القادمة من نظرية التصنيف (**category theory**)ـ الحقل الذي يدرس البُنى الرياضية والمفاهيم عالية التجريد ـ يُمكن استخدامها في تحليل الصورة إلى أجزائها

لمعرفة كيف تتلاءم المكونات المختلفة معاً. ويسمح ذلك الأمر، في سياق مجال تعلم الآلة (**machine learning**)، للآلة بأن "تفهم" ماهيّة الأجسام الموجودة في صورة ما، ومن ثمّ تتخذ قرارا "معقولا" بناءً على ذلك.

هذه قائمة قصيرة فقط، وفي الحقيقة فإنّ هناك العديد من الحقول الرياضية الأخرى وأيضاً علوم الحاسوب التي وجدت العديد من التطبيقات لها في مجال دراسة البيانات الضخمة. راقب هذا الفضاء! وأنا واثق أننا سنرى تطورات عظيمة في مجال الرياضيات الحسابية التطبيقية والبحتة نتيجة دراسة تلك التحديات.

عن المؤلف:



كريس بود

كريس بود **Chris Budd** بروفسور في الرياضيات التطبيقية في جامعة باث، ونائب رئيس معهد الرياضيات وتطبيقاتها، كما أنه استاذ للرياضيات في المعهد الملكي وزميل فخري للجمعية البريطانية للعلوم. وهو مهتم جزئياً بتطبيق الرياضيات على العالم الحقيقي، والمساهمة في جعل العامة يفهمون الرياضيات. شارك مع سانغوين في كتابة "**Mathematics Galore**!" الذي لاقى رواجاً كبيراً، ونشرته مطبعة جامعة اكسفورد.

- التاريخ: 2016-08-09
- التصنيف: أسئلة كُبرى

#المعلومات #البيانات الضخمة #البيانات

المصطلحات

- **تعليم الآلة (machine learning)**: تعلم الآلة هو أحد أنواع الذكاء الاصطناعي، يمكّن التطبيقات البرمجية من التنبؤ بنتائج أكثر دقة دون برمجتها بشكل صريح. ويتم ذلك عن طريق بناء خوارزميات تتلقى بيانات الإدخال وتستخدم التحليل الإحصائي للتنبؤ بقيمة المخرجات ضمن نطاق مقبول.
- **التنقيب ضمن البيانات (Data mining)**: هي عملية البحث ضمن مجموعة ضخمة من البيانات لتحديد الأنماط وإنشاء العلاقة فيما بينها ومن ثم تحليل هذه البيانات لحل مشاكل معينة.
- **البيانات الضخمة (Big data)**: مصطلح متطور يصف كمية ضخمة جداً من البيانات المُهيكلة وغير المُهيكلة، يمكن تحليلها حسابياً للحصول على الأساليب والنزعات والعلاقات والروابط، خصوصاً تلك التي تتعلق بالسلوك والتفاعل البشري.

المصادر

- plus magazine

المساهمون

- ترجمة
  - همام بيطار
- مُراجعة
  - أحمد ميمون الشاذلي
- تحرير
  - معاذ طلفاح
- نشر
  - سارة الراوي